



An unsupervised spatial-temporal-spectral fusion model for hyperspectral images with spectral upsampling residual network

Haoyang Yu, Xueting Wang, Hongyu Xin, Ke Zheng, Hongyan Zhang & Jiaochan Hu

To cite this article: Haoyang Yu, Xueting Wang, Hongyu Xin, Ke Zheng, Hongyan Zhang & Jiaochan Hu (2025) An unsupervised spatial-temporal-spectral fusion model for hyperspectral images with spectral upsampling residual network, International Journal of Remote Sensing, 46:22, 8460-8487, DOI: [10.1080/01431161.2025.2570551](https://doi.org/10.1080/01431161.2025.2570551)

To link to this article: <https://doi.org/10.1080/01431161.2025.2570551>



Published online: 27 Oct 2025.



Submit your article to this journal [↗](#)



Article views: 37



View related articles [↗](#)



View Crossmark data [↗](#)



ARTICLES



An unsupervised spatial-temporal-spectral fusion model for hyperspectral images with spectral upsampling residual network

Haoyang Yu^{a,b,c}, Xueting Wang^a, Hongyu Xin^a, Ke Zheng^d, Hongyan Zhang^{b,c} and Jiaochan Hu^a

^aCenter of Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian, China; ^bEngineering Research Center of Natural Resource Information Management and Digital Twin Engineering Software, Ministry of Education, Wuhan, China; ^cHubei Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences, Wuhan, China; ^dCollege of Geography and Environment, Liaocheng University, Liaocheng, China

ABSTRACT

Hyperspectral images (HSI) are renowned for their high spectral resolution and extensive wavelength coverage, but suffer from limited spatial and temporal resolution due to imaging sensor constraints. However, hyperspectral images with low spatial resolution and low temporal resolution are difficult to be applied to subsequent more advanced tasks such as object detection, classification, and anomaly detection. Physical constraints make it impossible for a single satellite sensor to acquire images that simultaneously have high resolution in time, space, and spectrum, so image fusion is the most efficient choice to achieve this goal. Spatial-temporal-spectral fusion (STSF) has the purpose of synthesizing different information which has the advantage in temporal, spatial, and spectral aspect respectively from multisource satellite data to reconstruct HSI with high spatial resolution and high temporal resolution. In order to address the problems that the linear relationship in current spectral reconstruction is difficult to accurately map the complex relationships among space, time and spectrum, and that deep convolutional neural networks are prone to overfitting, as well as to enhance the model's ability to extract spectral features, this paper designs an unsupervised STSF method. The proposed method has three stages: Stage 1 (Spatial-Spectral Downsampling) analyzes the spatial-spectral degradation of time 1 observed images; Stage 2 (Spectral Upsampling) develops a spectral upsampling network (with the shared spatial-spectral downsampling network) to upsample multispectral data to hyperspectral data; Stage 3 uses the trained network to upsample multispectral images of time 2 for high-spatial-resolution HSI. In order to verify the proposed method, it is compared with other state-of-the-art methods on simulated and real datasets. This proves the method's advantages in richer spatial-spectral details and more accurate reconstruction.

ARTICLE HISTORY

Received 12 March 2025

Accepted 28 September 2025

KEYWORDS

Spatial-temporal-spectral fusion; hyperspectral images; convolutional neural networks; unsupervised learning

1. Introduction

Hyperspectral images (HSI) have the advantages of high spectral resolution and wide wavelength coverage (Yu et al. 2024a). Due to inherent constraints of the imaging sensor, hyperspectral images often exhibit relatively limited spatial and temporal resolution, which hinders their capability in playing a role in the acquisition of fine surface information and the observation of continuity of time scales (Yang et al. 2023). The physical limitations of imaging sensors make it impossible for a single satellite sensor to obtain images with high resolution simultaneously in time, space and spectrum, so the fusion of remote sensing images from different satellite sensors is the only possibility to achieve this goal (Hou et al. 2025; Zhu et al. 2023, 2024). The goal of spatial-temporal-spectral fusion (STSF) is to synthesize the temporal, spatial, and spectral information from multisource remote sensing images to reconstruct HSI with high spatial and temporal resolution (W. Sun et al. 2020). In recent years, STSF technology has emerged as a key focus in multisource remote sensing data fusion research (W. Sun et al. 2023). At present, STSF methods are mainly divided into spatial-spectral fusion (SSF) methods, spatial-temporal fusion (STF) methods and STSF methods.

1.1. Spatial-spectral fusion method

Spatial-spectral fusion methods aim to generate images with high spatial-spectral resolution by integrating the spatial and spectral advantages of multisource remote sensing images (J. Li et al. 2023; Zheng, Khader, and Xiao 2022). SSF methods are mainly divided into three categories: the detail injection method, the model optimization method and the deep learning method (Jia et al. 2025; J. Li et al. 2022).

The method of detail injection is to separate the spectral information from the spatial information of the image, and then replace the spatial information in the hyperspectral information of low spatial resolution with the spatial information in the multispectral image of high spatial resolution, and finally obtain the hyperspectral image with high spatial resolution (Rahmani et al. 2010). Richard, Amin, and Menas (2001) first applied wavelet decomposition technology to this field, and band-by-band fusion of hyperspectral images was carried out using band grouping and completion methods. Based on wavelet decomposition, Y. Zhang and He (2007) adopted 3D wavelet transform technology to align input images by using spatial spectral resampling technology and then carried out wavelet coefficient fusion. Sylla et al. (2014) selected the most appropriate multispectral bands for each hyperspectral band for fusion through the calculation of correlation coefficients.

The model optimization method is to establish the degradation function model between the hyperspectral image with high spatial resolution, the multispectral image with high spatial resolution and the hyperspectral image with low spatial resolution (Qu, Qi, and Kwan 2018; Y. Zhang, De Backer, and Scheunders 2009). Aiming at the super resolution of hyperspectral images, Dian, Li, and Fang (2016) proposed a sparse representation based on non-local use of similar patterns and structures of low spatial resolution images to improve the efficiency of sparse solution. Dong et al. (2016) proposed structured sparse regular terms based on clustering to learn spatial correlation. Simões

et al. (2015) assumed that the target hyperspectral exists in a subspace of lower dimensions and solved the coefficient by applying the total variational regular term.

Deep learning methods use multi-layer neural networks to fit the correspondence between data in different images (Yan et al. 2025). Yu et al. (2024b) proposed an unsupervised fusion method for hyperspectral and multispectral images based on deep spectral-spatial cooperative constraints. Based on the unsupervised learning framework, the model uses the idea of degradation model to design a processing technology to eliminate the relative radiation differences between different sources of data, so as to achieve effective fusion of multi-source data. Qu, Qi, and Kwan (2018) assumed that the abundance of the input image obeys the Dirichlet distribution so that it satisfies the sum-one constraint. On this basis, Zheng et al. (2021) realized the adaptive learning of the degenerate model, thus getting rid of the dependence on the prior of the degenerate function. Yao et al. (2020) introduced the cross-modal cross-attention mechanism to realize the efficient use of information between multispectral images and hyperspectral images.

1.2. Spatial-temporal fusion method

Spatial-temporal fusion methods are to generate high spatial-temporal resolution images by integrating the high spatial but low temporal resolution images and the low spatial but high temporal resolution images. STF methods can be divided into the weight function method, the unmixing method, the Bayesian method and deep learning method (Wang and Atkinson 2018).

The weight function method achieved accurate image fusion output through linear combination of input image information (Feng et al. 2006). A spatial-temporal approach based on bilateral filters proposed by Huang et al. (2013a) took into account the temperature of ground objects in urban areas and used bilateral filters to determine the weights of adjacent pixels in STARFM to generate a high spatial-temporal resolution land surface temperature map. Wang et al. (2017) used the advanced region-to-point regression Kriging method to fuse MODIS and Landsat data and the method is devoted to solving the information of abrupt and heterogeneous landforms. Y. Sun, Hua, and Shi (2019) proposed a method of spatial-temporal fusion of remote sensing images by using linear injection model and local neighbourhood information.

The unmixing method assumes that the coarse pixels are mixed pixels and estimates the exact pixel value by unmixing the coarse pixels (B. Chen, Huang, and Xu 2015). Zurita-Milla, Clevers, and Schaepman (2008) introduced constraints in the linear decomposition to ensure that the value of the reflectivity change is positive and within an appropriate range. Maselli and Rembold (2002) explained the spatial variability of intra-class NDVI by introducing a locally calibrated multivariate regression model in non-mixing. Wu et al. (2015) proposed a spatiotemporal fusion method for adaptive window size selection, which selected the best window size and moving steps for coarse pixel decomposition to avoid the constant window and sensor differences of decomposition.

The Bayesian method transforms the space-time fusion problem into a probability distribution problem and obtains the image fusion result by maximizing the posterior probability density (J. Li et al. 2020). A. Li et al. (2013) used Bayesian maximum entropy to mix sea surface temperatures from MODIS images

and AMSR-E images. Huang et al. (2013b) proposed a unified Bayesian framework for spatial spectral fusion. Wei et al. (2017) proposed a spatiotemporal fusion model, which uses the maximum posterior probability to describe the inverse fusion problem.

With the development of deep learning, network models such as traditional convolutional neural network (CNN), generative adversarial network, and Transformer have been used to predict accurate Landsat images based on MODIS images (Y. Li et al. 2020). By improving the dictionary learning process and based on the single pair SPSTFM algorithm, D. Li et al. (2018) proposed an enhanced spatial-temporal fusion scheme based on a proposed fusion method with two expansion modes. Tan et al. (2018) proposed a new data fusion model, that is, deep convolutional spatiotemporal fusion network, which makes full use of convolutional neural networks to derive high spatiotemporal resolution images from remote sensing images with high spatiotemporal resolution and low spatiotemporal resolution.

1.3. Spatial-temporal-spectral fusion method

STSF methods aim to synthesize the spatial, temporal, and spectral information from multisource remote sensing images to reconstruct hyperspectral images with high spatial and temporal resolution. According to the stages and steps of fusion, STSF methods can be divided into the step-by-step fusion method and the integrated fusion method (W. Sun et al. 2023).

Step-by-step fusion divides the spatial-temporal-spectral fusion process into stages. A spatial temporal spectral mixing model is proposed by L. Zhang et al. (2016) to predict surface reflectance with high temporal, spatial and spectral resolution. Zhao and Huang (2017) proposed a hybrid spatial-spectral image fusion model, which integrates temporal, spatial and spectral information from Landsat, Hyperion and MODIS data. The model consists of three fusion modules: spatial-spectral, spatial-temporal and temporal-spectral fusion. Jiang, Shen, and Li (2022) proposed for the first time a heterogeneous integration framework based on a novel deep residual cyclically generated adversarial network, which includes a forward fusion part and a backward degenerate feedback part, to integrate spatial, temporal and spectral information of multi-source remote sensing images.

Unlike step-by-step fusion, which separates the fusion process into distinct stages, integrated fusion adopts a unified modelling approach to simultaneously achieve high spatial, temporal, and spectral resolution (Huang et al. 2013). A novel unsupervised three-dimensional tensor space decomposition network is proposed by W. Sun et al. (2023). Combined with 3-D tensor space decomposition theory, the method uses 3-D hyperspectral/multispectral tensor space extraction network to predict the low spatial resolution hyperspectral tensor space features that are missing at other times. X. Chen et al. (2023) proposed a progressive STSF network (PSSTFN), which integrates spatial-spectral fusion and spectral-temporal fusion into a unified end-to-end STSF framework to extract hierarchical features of different sensitivity fields, as well as feature extraction and insertion for spatial-temporal spectrum fusion. Zhou et al. (2022) proposed a new STSF generalized linear spectrum mixing model. The generalized linear spectral mixing model is introduced into the STSF problem, and the time variation of the image at different time is transferred to the end element and abundance matrix of the image for estimation.

At present, there are some STSF methods, but these methods still have some problems. Firstly, most existing spatial-temporal-spectral fusion methods are restricted to the assumption of linear relationship among temporal, spatial, and spectral aspects. However, due to nonlinear optical effects such as atmospheric scattering and absorption, nonlinear effects of mixed pixels and other factors, the relationship between multispectral and hyperspectral is more complex, and it is difficult to accurately reconstruct hyperspectral data through a simple linear mixing model. In order to solve the nonlinear problem in spectral reconstruction, the depth residual network based on convolutional neural network is used in this paper to effectively reconstruct spectral information. In addition, most STSF methods use MODIS and Landsat data, where hyperspectral MODIS has higher temporal resolution, in contrast to domestic hyperspectral data ZY1-02D has lower temporal resolution. This makes it difficult for the existing STSF methods to be directly applied to the domestic hyperspectral data ZY1-02D. Therefore, this paper designed the STSF method for low temporal resolution hyperspectral data and high temporal resolution multispectral data. Finally, there is not a sufficient number of high spatial-temporal-spectral resolution images as training data for supervised training leading serious limitations on supervised approaches. Thus, this paper proposes an unsupervised spatial-temporal-spectral fusion model for hyperspectral images. To clearly present the differences between this study and existing methods, the core innovations are summarized as follows:

- (1) This paper proposed an efficient unsupervised STSF method, which completely relies on low spectral resolution hyperspectral images and high spatial resolution multispectral images at time 1 and high spatial resolution multispectral images at time 2, and does not require the participation of other additional data.
- (2) In this paper, a spectral upsampling network with adaptive spectral channel is designed to enhance the multi-layer feature extraction ability of the network, and a global shared spatial-spectral downsampling network is used to accurately represent the nonlinear constraints of spatial-spectral degradation.
- (3) This method is practical and generalizable. The proposed method is compared with other state-of-the-art methods based on simulated and real datasets (the inter-temporal LN-02T dataset), and it is demonstrated that the proposed method has richer spatial spectral details and more accurate reconstruction results.

2. Proposed approach

In this paper, $\mathbf{X}_i \in \mathbb{R}^{W \times H \times c}$ represents the high spatial resolution multispectral images (HRMSI) at T_i , $\mathbf{Y}_i \in \mathbb{R}^{w \times h \times C}$ represents the low spectral resolution hyperspectral images (LRHSI) at T_i , $\mathbf{W}_i \in \mathbb{R}^{w \times h \times c}$ represents the low spatial resolution multispectral images (LRMSI) at T_i , $\mathbf{Z}_i \in \mathbb{R}^{W \times H \times C}$ represents the high spectral resolution hyperspectral images (HRHSI) at T_i . $w \times h$ and $W \times H$ are the size of spatial dimension of the images this paper mentioned ($W > w, H > h$). Respectively, c and C are the number of spectral bands of the images this paper mentioned ($C > c$). In order to distinguish the estimated and real images in the model, \mathbf{F}_1 , \mathbf{F}_2 and \mathbf{F}_3 are used to represent the input real HRMSI at T_1 , LRHSI at T_1 and HRMSI at T_2 .

The model proposed in this paper, as shown in [Figure 1](#), is an unsupervised spatial-temporal-spectral fusion model for hyperspectral images with spectral upsampling residual network (UFSURN). It combines the information from a set of HRMSI $\mathbf{F}_1 \in \mathbb{R}^{W \times H \times c}$ and LRHSI $\mathbf{F}_2 \in \mathbb{R}^{w \times h \times C}$ at T_1 and HRMSI $\mathbf{F}_3 \in \mathbb{R}^{W \times H \times c}$ at T_2 to obtain HRHSI $\mathbf{Z}_2 \in \mathbb{R}^{W \times H \times C}$ at T_2 . The framework of the proposed UFSURN is shown in [Figure 1](#), including three stages of the Spatial-Spectral downsampling, Spectral Upsampling and Reconstruction. In the first stage, this paper uses the LRHSI and HRMSI of time 1 to establish the spectral downsampling network and the spatial downsampling network, which are used to obtain the degradation information required later, including the parameters of the two LRMSI, the trained spectral downsampling network and the spatial downsampling network. In the second stage, this paper uses the two degeneration networks trained in the first stage, HRMSI of time 2, and LRMSI according to the LRHSI of time 1, to train the spectral upsampling network. In the third stage, the trained spectral upsampling network and HRMSI of time 2 are used to reconstruct the HRHSI of time 2. It is worth noting that the whole process is performed on the observed LRHSI and HRMSI of time 1 and the HRMSI of time 2 without the participation of any additional data.

2.1. Spatial-spectral downsampling

The traditional CNN can only process local spectral information stage by stage in spectral super-resolution, and it is difficult to model the global correlation between different bands. The spatial-spectral downsampling network of UFSURN employs a globally shared convolutional neural network to simultaneously model spatial degradation (such as Gaussian blur point spread function (PSF)) and spectral degradation (such as spectral response function (SRF)) through shared parameters. This design enables the network to capture the global nonlinear spatial-spectral relationship across stages, such as sharing

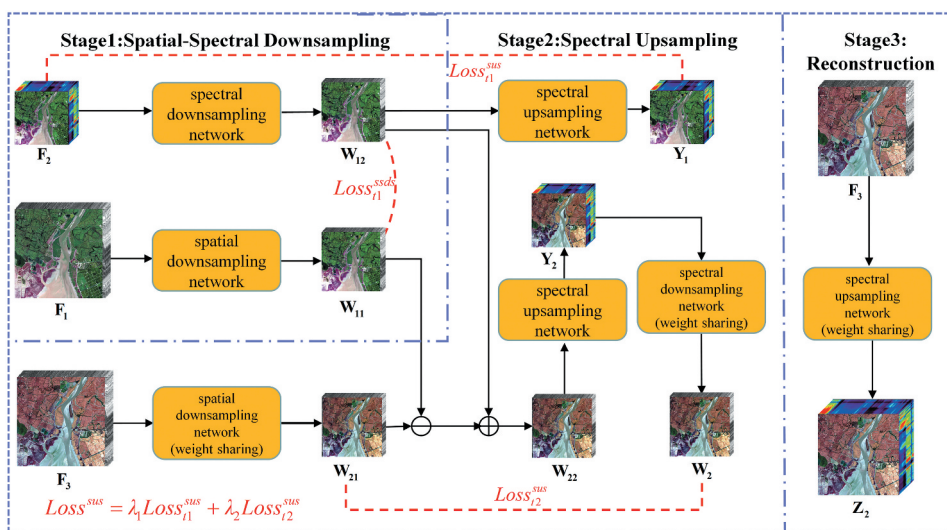


Figure 1. Overview of the proposed UFSURN.

convolution kernel parameters in the degradation model, so as to more accurately describe the joint degradation process of multi-source data. The spatial downsampling network and spectral downsampling network are introduced in the following, respectively.

The spatial downsampling network's objective is to downsample the input HRMSI to LRMSI on spatial aspect. In the spatial downsampling network, the PSF is used to realize the conversion of HRMSI to LRMSI. PSF describes the imaging diffusion characteristics of an optical system for an ideal point light source and reflects the degree of blurring of image details by the optical system.

In the specific design of the spatial downsampling network, the convolution kernel of the convolutional layer is initialized in the form of a Gaussian function. In the initial stage, a large convolution kernel of size $n \times n$, where $n = W/w = H/h$ in order to reduce the spatial scale from W to w , is used to capture the large-scale spatial diffusion information of the image and realize preliminary downsampling. With the increase of the number of network layers, the size of convolution kernel is gradually reduced into 1×1 which is used to fine-tune local features and simulate the change of PSF in small-scale space. At the same time, the introduction of the pooling layer can further reduce the spatial resolution of the image, which is combined with the diffusion effect of the Gaussian function to achieve larger downsampling.

The above process can be abstractly represented using mathematical notation as follows:

$$\mathbf{W}_{11} = f_{11}(\mathbf{F}_1; \theta_1) \quad (1)$$

where f_{11} denotes the spatial downsampling network and θ_1 denotes the learnable parameters of spatial downsampling. Here, \mathbf{W}_{11} specifically refers to the LRMSI generated by inputting the HRMSI at time 1 into the spatial downsampling network.

The spectral downsampling network's objective is to downsample the input LRHSI to LRMSI on spectral aspect. In the spectral downsampling network, the SRF is used to simulate spectral downsampling. The SRF describes the response characteristics of an imaging sensor to different wavelengths of light and determines how the sensor converts continuous spectral information into discrete spectral channels. In practice, the spectral responses of different sensors vary due to their physical characteristics and design limitations. By using SRF, the LRHSI can be accurately downsampled to the LRMSI in the spectral dimension to match the spectral properties of the actual sensor.

The CNN has powerful feature learning capabilities. In CNN, the convolution kernel can simulate the role of SRF. Specifically, the weight parameters of the convolution kernel can be learned to fit a response curve close to the actual sensor SRF. The shape of the convolution kernel is set to a known spectral response function to model a reduction in the number of spectra, so that each spectral band in the multispectral image captures features from all bands related to itself in the hyperspectral image. To ensure that the input and output space sizes remain the same, the stride parameter is set to 1 and the convolution kernel for size is set to 1×1 . The activation function selected the nonlinear function ReLU to increase the nonlinear expression ability of the network and better simulate the complex characteristics of the spectral response function. Through multiple

layers of convolution and activation operations, the feature extraction and downsampling of hyperspectral data are gradually realized.

This can be abstractly represented using mathematical notation as follows:

$$\mathbf{W}_{12} = f_{12}(\mathbf{F}_2; \theta_2) \quad (2)$$

where f_{12} denotes the spatial downsampling network and θ_2 denotes the learnable parameters of spatial downsampling.

The above structure pertains to the proposed spatial-spectral downsampling network, whose core goal is to learn the spatial degradation (modelled by PSF) and spectral degradation (modelled by SRF) processes of remote sensing images, and accurately estimate the degradation parameters for subsequent spectral upsampling. To ensure the reliability of parameter estimation, the selection of loss function must align with the characteristics of ZY1-02D satellite data and the task objectives of this stage.

In this stage, the input LRHSI and HRMSI may contain local abnormal pixels caused by two factors: (1) sensor noise of the ZY1-02D hyperspectral payload (e.g. spectral noise under low illumination conditions); (2) residual atmospheric scattering interference after preprocessing, which distorts individual pixel values. For the loss function selection, we compare the L-1 norm (Mean Absolute Error (MAE)) and L-2 norm (Mean Squared Error (MSE)) losses, and finally choose the L-1 norm loss function for the following reasons:

The L-2 norm loss calculates the average of squared differences between predicted and true values. Due to the square operation, the loss term of abnormal pixels (with large errors) will be amplified exponentially. For Equation 3 – which aims to optimize the spatial-spectral downsampling network to fit the true degradation process – such amplified loss will dominate the training process, forcing the network to adjust PSF and SRF parameters to fit abnormal pixels. This will deviate the learned degradation model from the actual imaging physics, making it impossible to provide accurate degradation information for the subsequent spectral upsampling network.

The L-1 norm loss calculates the average of absolute differences between predicted and true values. Its linear punishment mechanism for errors avoids excessive emphasis on abnormal pixels, and the loss value can more stably reflect the fitting degree of the network to most normal data. For Equation 3 this ensures that the learned PSF and SRF parameters conform to the real spatial diffusion and spectral response characteristics, laying a solid foundation for accurate spectral upsampling network.

The L-1 norm loss function for the spatial-spectral downsampling network is abstractly represented using mathematical notation as follows:

$$Loss_{t1}^{ssds} = \frac{1}{whc} \|\mathbf{W}_{11} - \mathbf{W}_{12}\|_1 \quad (3)$$

2.2. Spectral upsampling

In this stage, this paper uses the two degeneration networks trained in the first stage, LRMSI of time 2, and LRMSI according to the LRHSI of time 1, to train the spectral upsampling network. In the spectral upsampling network, this paper uses the SUREsNet (for a detailed introduction to SUREsNet, see subsection 2.3) to achieve spectral upsampling. This can be abstractly represented using mathematical notation as follows:

$$\mathbf{Z}_2 = f_2(\mathbf{F}_3; \theta_3) \quad (4)$$

where f_2 denotes the spectral upsampling network and θ_3 denotes the learnable parameters of spectral upsampling network.

To train the spectral upsampling network, firstly, the LRMSI $\mathbf{W}_{21} \in \mathbb{R}^{w \times h \times c}$ at T_2 is obtained by using the trained spatial downsampling network with the HRMSI $\mathbf{F}_3 \in \mathbb{R}^{W \times H \times c}$ at T_2 . In short time intervals, the difference between the two observations of the same scene or object is stable without drastic changes in its properties. That means the difference between LRMSI $\mathbf{W}_{11} \in \mathbb{R}^{w \times h \times c}$ at T_1 and LRMSI $\mathbf{W}_{12} \in \mathbb{R}^{w \times h \times c}$ at T_1 and the difference between LRMSI $\mathbf{W}_{21} \in \mathbb{R}^{w \times h \times c}$ at T_2 and LRMSI $\mathbf{W}_{22} \in \mathbb{R}^{w \times h \times c}$ at T_2 should be stable. And end up with $\mathbf{W}_{22} \in \mathbb{R}^{w \times h \times c}$ based on this equation. This can be abstractly represented using mathematical notation as follows:

$$\mathbf{W}_{11} - \mathbf{W}_{21} = \mathbf{W}_{12} - \mathbf{W}_{22} \quad (5)$$

$$\mathbf{W}_{22} = \mathbf{W}_{21} - \mathbf{W}_{11} + \mathbf{W}_{12} = f_{11}(\mathbf{F}_3; \theta_1) - \mathbf{W}_{11} + \mathbf{W}_{12} \quad (6)$$

Finally, LRMSI $\mathbf{W}_{22} \in \mathbb{R}^{w \times h \times c}$ goes through spectral upsampling network to obtain the LRHSI $\mathbf{Y}_2 \in \mathbb{R}^{w \times h \times C}$. And LRMSI $\mathbf{W}_2 \in \mathbb{R}^{w \times h \times c}$ from LRHSI $\mathbf{Y}_2 \in \mathbb{R}^{w \times h \times C}$ at T_2 was obtained through the shares trained spectral downsampling network. This can be abstractly represented using mathematical notation as follows:

$$\mathbf{W}_2 = f_{12}(\mathbf{Y}_2; \theta_2) = f_{12}(f_2(\mathbf{W}_{22}; \theta_3); \theta_2) \quad (7)$$

Here, \mathbf{W}_2 is the output result obtained via the following process: first, the known HRMSI at time 2 spatially downsamples and processes temporal difference to generate the LRMSI; then, the LRMSI undergoes a two-step processing. Specifically, in the first step, the LRMSI is fed into the spectral upsampling network (SUNet) to realize spectral dimension expansion (from multispectral to hyperspectral); in the second step, the expanded result is input into a pre-trained spectral downsampling network (shared with the spatial-spectral downsampling network in the first stage) for spectral downsampling.

At the same time, this paper takes LRMSI $\mathbf{W}_{12} \in \mathbb{R}^{w \times h \times c}$ into the spectral upsampling network which is training to get the estimated LRHSI $\mathbf{Y}_1 \in \mathbb{R}^{w \times h \times C}$. This can be abstractly represented using mathematical notation as follows:

$$\mathbf{Y}_1 = f_2(\mathbf{W}_{12}; \theta_3) \quad (8)$$

This paper designs a set of weighted loss functions (Equation 9) to constrain the optimization of the spectral upsampling network, with the core goal of reconstructing LRHSI that is consistent with the true spectral trend from LRMSI (without relying on HRHSI ground truth in the unsupervised scenario). The selection of the L-1 norm (MAE) loss function instead of the L-2 norm (MSE) loss is determined by the unsupervised training characteristics and spectral reconstruction requirements of this stage:

First, in the unsupervised framework of this paper, the spectral upsampling network (SUNet) cannot obtain HRHSI ground truth for supervision, and can only rely on the consistency between the reconstructed LRHSI (via the spectral upsampling network) and the input LRHSI for optimization. At this time, the tiny noise in the LRHSI will be amplified by the square operation of the L-2 norm loss, leading to two problems: (1) the model overfits to noise, resulting in spectral jitter of the reconstructed LRHSI (e.g. abnormal

fluctuations in the red edge band of vegetation); (2) the gradient of the L-2 norm loss is proportional to the error (gradient = $2 \times \text{error}$), which will cause the gradient to explode when the error is large (e.g. initial training stage), making SUREsNet difficult to converge. Second, the L-1 norm loss avoids the above defects: (1) its linear punishment mechanism for errors reduces the impact of tiny noise on training, allowing the model to focus on the overall spectral trend reconstruction (e.g. the spectral reflection peak of Suaeda salsa in the near-infrared band), which is consistent with the goal of Equation 9 to ensure spectral fidelity; (2) the gradient of the L-1 norm loss is a constant (1 or -1 , except when the error is 0), which maintains stable gradient propagation during SUREsNet training, especially in the deep residual structure, avoiding gradient vanishing or explosion caused by the L-2 norm loss.

In addition, the weighted design of Equation 9 further coordinates the constraints of multiple terms, and the L-1 norm loss ensures that the weight adjustment of each term is not disturbed by abnormal values, thereby improving the robustness of the model and the quality of spectral reconstruction. λ_1 is the weight of the temporal consistency constraint, which balances the smoothness of time-series data and avoids overfitting to individual frames; λ_2 is the weight of the spectral fidelity constraint, which regulates the trade-off between spectral reconstruction accuracy and noise robustness. The specific form of the loss function is as follows:

$$Loss^{sus} = \lambda_1 Loss_{t1}^{sus} + \lambda_2 Loss_{t2}^{sus} = \lambda_1 \frac{1}{whc} \|\mathbf{W}_2 - \mathbf{W}_{21}\|_1 + \lambda_2 \frac{1}{whc} \|\mathbf{F}_2 - \mathbf{Y}_1\|_1 \quad (9)$$

2.3. SUREsNet block

Spectral upsampling, as a key step in the conversion from MSI to HSI, faces a core challenge of achieving precise spectral dimension expansion under limited training data conditions. Traditional convolutional neural networks (CNNs) have three significant limitations in this task: Firstly, information passing through successive layers is prone to cause gradient vanishing or explosion problems, affecting the efficiency of deep feature learning; Secondly, existing residual network modules (such as the basic residual blocks in ResNet) generally adopt a fixed channel design, making it difficult to adapt to the significant channel expansion requirements from multi-spectral (8 bands) to hyperspectral (79 bands), and direct expansion would lead to feature disorder; Thirdly, most spectral super-resolution methods rely on the supervised learning paradigm and require a large amount of high-resolution high-spectral true value data (HRHSI) for training, while actual data sources such as the ZY1-02D satellite have problems of low time resolution (>60 days) and scarce true value samples, severely limiting the application scenarios of the supervised methods.

To address these limitations, this paper proposes the Spectral Upsampling Residual Network (SUREsNet), whose design motivation is clear: through a ‘multi-stage convolution + adjacent layer shortcut + adaptive channels’ collaborative architecture, efficient spectral feature expansion and retention can be achieved in an unsupervised scenario. This module serves as the core component of the UFSURN model, with the input being a multispectral image, and the output being the corresponding hyperspectral image,

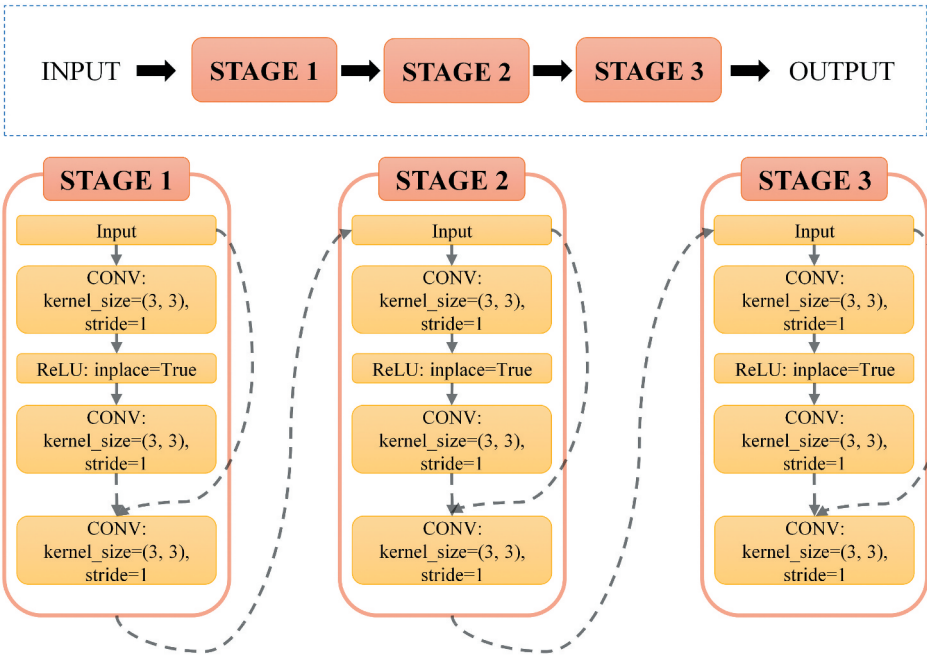


Figure 2. Structure of the SUREsNet block.

providing the network foundation for the third stage’s spectral resolution reconstruction.

2.3.1. Overall architecture of the module

SUREsNet adopts a three-stage residual structure (as shown in Figure 2), using a ‘small-step progressive’ channel expansion strategy to avoid spectral information loss caused by a single large-scale expansion. Each stage consists of three layers of 3×3 convolutions, ReLU activation function, and short connections between adjacent layers within the stage. The specific structure is as follows:

- (1) Stage 1 (Channel $8 \rightarrow 24$): Taking the 8-channel feature of MSI as the input, the first layer of 3×3 convolution completes the initial feature extraction, and the ReLU activation function enhances the nonlinear expression ability; the second layer of 3×3 convolution completes further feature extraction, and the output features are added to the input of the first phase to form a residual connection, which serves as the input of the third layer; the third layer of 3×3 convolution expands the number of channels to 24, achieving the initial upsampling in the spectral dimension.
- (2) Stage 2 (Channel $24 \rightarrow 48$): For the 24-channel features output in the first stage, the ‘convolution \rightarrow ReLU \rightarrow convolution \rightarrow convolution’ operation is repeated to expand the number of channels to 48. Similar to the intra-layer short-circuit design, the output of the second layer convolution is fused with the input features of the Stage 2 to ensure the stable transmission of spectral information.

- (3) Stage 3 (Channel 48→79): As the final spectral expansion stage, through three layers of 3×3 convolution, the number of channels is expanded from 48 to 79 (consistent with the target high-resolution hyperspectral image band number), and after short-circuiting the output of the second layer of convolution with the input features of Stage 3 through the third layer of 3×3 convolution, the final HSI features are obtained.

From a mathematical perspective, the feature transformation process of SUREsNet can be expressed as:

$$\mathbf{Y}_1 = \text{Conv}_{3 \times 3, 8 \rightarrow 24}(\text{Conv}_{3 \times 3}(\text{Relu}(\text{Conv}_{3 \times 3} \mathbf{X}_{\text{INPUT}})) + \mathbf{X}_{\text{INPUT}}) \quad (10)$$

$$\mathbf{Y}_2 = \text{Conv}_{3 \times 3, 24 \rightarrow 48}(\text{Conv}_{3 \times 3}(\text{Relu}(\text{Conv}_{3 \times 3} \mathbf{Y}_1)) + \mathbf{Y}_1) \quad (11)$$

$$\mathbf{Y}_{\text{OUTPUT}} = \text{Conv}_{3 \times 3, 48 \rightarrow 79}(\text{Conv}_{3 \times 3}(\text{Relu}(\text{Conv}_{3 \times 3} \mathbf{Y}_2)) + \mathbf{Y}_2) \quad (12)$$

Among them, $\mathbf{X}_{\text{INPUT}}$ represents the input multispectral features, while \mathbf{Y}_1 and \mathbf{Y}_2 are the output features of Stage 1 and Stage 2, respectively, $\mathbf{Y}_{\text{OUTPUT}}$ represents the output hyperspectral features, and $\text{Conv}_{k \times k, C_{in} \rightarrow C_{out}}$ indicates a convolution operation with a kernel size of k , input channels of C_{in} , and output channels of C_{out} .

2.3.2. Key innovative design

Compared with the traditional residual module, the innovation of SUREsNet is demonstrated through the following three aspects:

- (1) Adaptive channel expansion mechanism: Abandoning the fixed channel design of traditional residual blocks, adopt a progressive expansion strategy of '8→24→48→79', with the expansion ratio controlled at 2-3 times per stage to avoid feature disorder caused by a large-scale expansion in a single step. This design not only conforms to the characteristic of 'strong correlation between bands' of high-spectrum data but also reduces the computational complexity.
- (2) Short-circuit optimization within stages: The 'adjacent layers within the stage short-circuiting' design is adopted – the output of the second layer of convolution in each Stage is directly added to the input features of that Stage, rather than being connected across stages. This design not only retains the advantage of traditional residual networks in alleviating the problem of gradient disappearance but also through multi-stage splitting, enables the spectral features to be fully learned in each expansion step, reducing the loss of band-related information.
- (3) Unsupervised scene adaptability: The training of SUREsNet only relies on LRHSI and HRMSI data at time 1 and HRMSI data at time 2. It achieves unsupervised constraints through the L_1 norm loss defined in Section 2.2, without the need for HRHSI true sample values. Compared with the strong dependence of supervised methods on labelled data, this design perfectly adapts to the actual scenario of scarce high-spectrum training data.

2.3.3. The connection with the overall model

SUResNet, as the core intermediate module of the UFSURN model for achieving spectral-spatial joint optimization, plays a crucial role in connecting the entire process of model training and inference. It achieves the connection between the second-stage training and the third-stage inference through 'cross-stage parameter transfer + data interface adaptation'.

During the second-stage training of the model, SUResNet undertakes the core task of learning the spectral mapping rules. In this stage, SUResNet optimizes the network parameters using the L_1 norm loss function, focusing on learning the nonlinear mapping relationship between multi-spectral bands and hyperspectral bands, especially the spectral response matching rules of 8-band MSI and 79-band HSI in the ZY1-02D satellite data. After this stage of training is completed, the convolution kernel parameters and residual connection weights of SUResNet are fixed, providing stable spectral conversion capabilities for the third-stage inference task.

In the third-stage inference process, SUResNet realizes the spectral upsampling function by adapting to the real-world data interface. This stage uses HRMSI at time 2 as the input for pre-training SUResNet, and completes the spectral expansion from 8 bands to 79 bands through a three-stage residual structure, ultimately generating HRHSI at time 2. This design ingeniously utilizes the spectral mapping knowledge learned by SUResNet during the training stage, avoiding the error accumulation problem caused by re-training the spectral module in traditional methods.

3. Experimental configuration and results

3.1. Experimental settings

3.1.1. Operational configuration

The data used in this paper are all processed with ENVI 5.6 including registration, radiometric calibration, atmospheric correction, reprojection and other data processing operations. Ultimately, all experiments are conducted on a Window11 computer with 16GB of RAM Intel (R) Core (Rahmani et al.) i7-8700K CPU memory and a single-GPU NVIDIA GeForce RTX 2080.

In spatial-spectral downsampling stage, this paper sets the learning rate to 0.001 and the number of iterations to 6000, where 3000 is the number of iterations for the initial learning and 3000 is the number of iterations for the learning rate to decay linearly to zero. In spectral upsampling stage, this paper sets the learning rate to $4e-3$ and the number of iterations to 14,000, where 7000 is the number of iterations for the initial learning and 7000 is the number of iterations for the learning rate to decay linearly to zero. We tested typical combinations ($\lambda_1 = 0.5/1.0/1.5$; $\lambda_2 = 0.6/0.8/1.2$) and excluded extreme values based on performance trends (e.g. $\lambda_1 < 0.5$ led to significant temporal inconsistency). Our selected values ($\lambda_1 = 1.0$, $\lambda_2 = 0.8$) were determined by balancing key metrics, ensuring basic reproducibility.

3.1.2. Datasets

The ZY1-02D satellite is a remote sensing satellite of China, which was successfully launched on 12 September 2019. ZY1-02D single satellite orbit return cycle is 55 days,

due to width, cloud cover and other factors, the actual data available time resolution is generally higher than 60 days on the HSI, leading to the lack of hyperspectral training data.

Thus, the ZY1-02D data is more suitable for unsupervised framework. ZY1-02D is equipped with both hyperspectral and multispectral payloads, providing 30 m spatial resolution hyperspectral data (166 bands) and 10 m spatial resolution multispectral data (9 bands, including coastal blue, yellow, and red edge bands). This dual-payload feature enables it to provide complete ‘low spatial-high spectral’ and ‘high spatial-low spectral’ data pairs, which fully meet the core data requirements of spatial-temporal-spectral fusion. The LN-02T dataset based on the LN-02 dataset located in Liaohe estuary, Panjin City, Liaoning Province, China, consists of two sets of paired hyperspectral and multi-spectral images at similar times to ensure region integrity to reflect the dual temporal transformation. The details of the dataset are shown in Table 1.

The LN-02T dataset contains typical features such as intertidal muds, *Phragmites australis*, *Suaeda salsa*, and aquaculture ponds. As a dynamic sedimentary environment, intertidal muds’ spectral characteristics are significantly affected by sediment grain size, moisture content and algal cover. These factors lead to large fluctuations in reflectance across different phases, which can effectively test the model’s ability to capture spectral variations caused by dynamic environmental changes – a key challenge for STSF methods. Paddy fields and *Suaeda salsa* have significant seasonal spectral variation. For example, *Phragmites australis* shows increased reflectance in the near-infrared band during the growing season (May–September) and decreased reflectance in the senescent season (October–November); *Suaeda salsa* changes from green (growing period) to red (mature period), with significant shifts in the red edge band. Such variations allow the model’s capability to track vegetation phenological transitions and associated spectral changes to be fully validated. Aquaculture ponds are affected by water quality (e.g. suspended solids,

Table 1. Information of experiment data.

Datasets		Sensor	Spatial resolution (m)	Spectral resolution (nm)	Acquisition times (Day Month Years)	
					Input data	Target data
LN-02T	Real Dataset-1	HSI (ZY1-02D)	30	400–1000	19 March 2022 (LRHSI&HRMSI)	13 May 2022 (HRHSI)
		MSI (ZY1-02D)	10	400–1000	13 May 2022 (HRMSI)	
	Real Dataset-2	HSI (ZY1-02D)	30	400–1000	6 September 2022 (LRHSI&HRMSI)	22 October 2022 (HRHSI)
		MSI (ZY1-02D)	10	400–1000	22 October 2022 (HRMSI)	
	Simulated Datasets-1	HSI (ZY1-02D)	30	400–1000	19 March 2022 (Simulated HSI&Simulated MSI)	13 May 2022 (LRHSI)
		MSI (ZY1-02D)	10	400–1000	13 May 2022 (Simulated MSI)	
	Simulated Datasets-2	HSI (ZY1-02D)	30	400–1000	6 September 2022 (Simulated HSI&Simulated MSI)	22 October 2022 (LRHSI)
		MSI (ZY1-02D)	10	400–1000	22 October 2022 (Simulated MSI)	

chlorophyll concentration) and anthropogenic management (e.g. water changing, bait casting), the spectra show strong spatial-temporal heterogeneity. This heterogeneity enables the verification of the model's adaptability to spectral disturbances caused by combined natural and human-induced factors.

As shown in Figure 3, the Real Dataset is collected from hyperspectral payload and multispectral payload of ZY1-02D including Real Dataset-01 of MSI and HSI on 19 March 2022 and 13 May 2022 and Real Dataset-02 of MSI and HSI on 6 September 2022 and 22 October 2022. The Real Dataset cropped the 532×532 pixels (hyperspectral image) region to ensure region integrity to reflect the dual temporal transformation. This paper selected 79 bands from HSI corresponding to the 8 bands of MSI. The dimension of HSI is (532, 532, 79) and the dimension of MSI is (1596, 1596, 8) in Real Dataset. The time intervals of the two datasets are 55 days (from 19 March to 13 May) and 46 days (from 6 September to 22 October), respectively, covering vegetation growth cycles (such as paddy fields from bud to heading) and phenological changes (such as vegetation wilting in autumn). Its two simulated subsets cover time intervals of 55 days (19 March to 13 May, involving paddy growth from bud to heading) and 46 days (6 September to 22 October, involving autumn vegetation wilting), and include typical features with distinct spectral characteristics (intertidal muds, *Phragmites australis*, *Suaeda salsa*, aquaculture ponds), which can fully simulate complex surface scenarios.

In order to prove the generalization of the model proposed in this paper, the experiment data includes simulated data and real data both based on LN-02T.

For the Real Dataset-1, the task is to use the HRMSI with dimensions of (1596, 1596, 8) of 19 March 2022 and 13 May 2022 and the LRHSI with dimensions of (532, 532, 79) of 19 March 2022 to predict HRHSI image with dimension of (1596, 1596, 79) of 13 May 2022. For the Real Dataset-2, the task is to use the HRMSI with dimension of (1596, 1596, 8) of

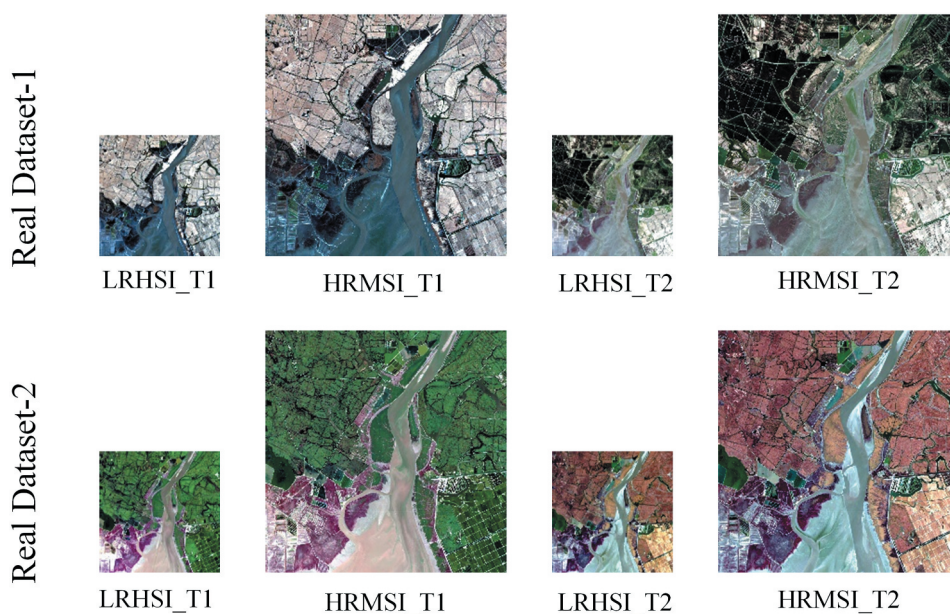


Figure 3. Real hyperspectral and multispectral dataset from ZY1-02D satellite (LN-02T dataset).

6 September 2022 and 22 October 2022 and the LRHSI with dimension of (532, 532, 79) of 6 September 2022 to predict HRHSI image with dimension of (1596, 1596, 79) of 22 October 2022.

For the Simulated Dataset, as shown in Figure 4, this paper uses the multispectral spectral response function of ZY1-02D to spectrally downsampled LRHSI with dimension of (532, 532, 79) to obtain simulated LRMSI with dimension of (532, 532, 8). At the same time, apply Gaussian filtering and spatially downsampled HSI with dimension of (532, 532, 79) to obtain simulated HSI with dimension of (133, 133, 79). The above downsampling operations accurately reproduce the imaging degradation mechanism of the ZY1-02D sensor. The task is to use the simulated MSI at T_1 and T_2 , and the simulated HSI at T_1 to predict HSI with dimension of (532, 532, 79) at T_2 . The HSI acquisition dates of Simulated Dataset-1 are 19 March 2022 and 13 May 2022. The HSI acquisition dates of Simulated Dataset-2 are 6 September 2022 and 22 October 2022.

3.1.3. Comparison experiment

Since most current STSF methods deal with hyperspectral MODIS with higher temporal resolution than multispectral Landsat, they are not suitable for hyperspectral ZY1-02D with lower temporal resolution than multispectral Sentinel-2. To verify the superiority of the UFSURN, this paper proposed, the spectral super-resolution method EDCSTFN (Tan et al. 2019) and the spatial-spectral fusion methods UDCNN (Yu et al. 2023) and UMC2FF (J. Li et al. 2023) are selected for comparison. EDCSTFN is a classic spectral super-resolution method widely used in hyperspectral image reconstruction. It focuses on modelling spectral feature mapping, which can effectively verify the superiority of our proposed SUREsNet (spectral upsampling residual network) in adaptive spectral

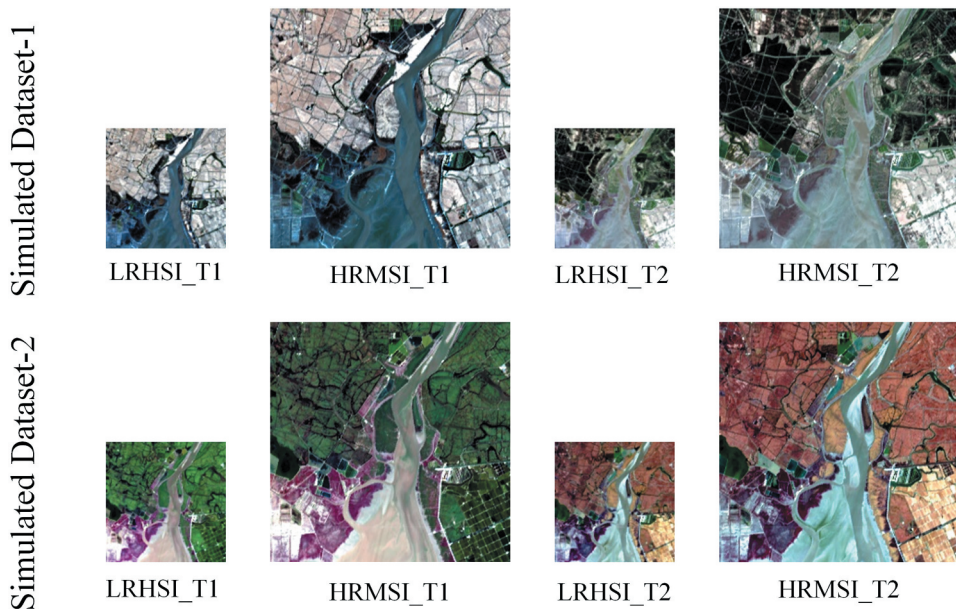


Figure 4. Simulated hyperspectral and multispectral dataset based on ZY1-02D spectral response function and Gaussian filtering (LN-02T dataset).

upsampling. UDCNN is a typical unsupervised spatial-spectral fusion method, which is highly consistent with the unsupervised framework of our UFSURN. Comparing with UDCNN can directly reflect the improvement of our global shared spatial-spectral downsampling network in handling nonlinear spatial-spectral degradation relationships. Both methods serve as targeted baselines to highlight the core innovations of UFSURN. In order to verify the superiority of the proposed method, two types of ablation experiments are performed on the core structure of the proposed method UFSURN, and the necessity of the optimal structure is verified by comparing different architecture configurations. The one ablation experiment named UFSURN-LS is designed to verify the importance of the nonlinear transformation for capturing the complex spectral-spatial mapping relationship, where the spatial spectral descent model was replaced by a linear structure (1-layer Conv2d and 1-layer ReLU). The other one ablation experiment named UFSURN-SL verified the necessity of the deep network to extract multi-level spectral features, and the neural network in the spectrum ascent model was replaced by a single-layer network (1-layer Conv2d and 1-layer ReLU).

3.1.4. Quality metrics for Simulated data

To perform the objective fusion quality assessment for the simulated data, this paper adopts five common metrics for quantitative assessment, including root mean square error (RMSE), mean peak signal-to-noise ratio (MPSNR), spectral angle mapping (SAM), structural similarity (SSIM) and relative global dimensional error in synthesis (ERGAS).

3.1.4.1. RMSE. Root mean square error (often abbreviated RMSE) is a measure of the size of a prediction model's prediction error, which is the square root of the sum of squares of the difference between the actual observed value and the predicted value of the model. Specifically, RMSE is calculated by the following formula:

$$RMSE(A, B) = \sqrt{\frac{\|A - B\|_F^2}{N}} \quad (13)$$

where N is the number of observed values, A is the observed value, and B is the predicted value of the model. This formula averages the square of the difference between the actual observed value and the predicted value of the model, and then takes the square root to get a global measure of error.

The smaller the value of RMSE, the higher the prediction accuracy of the model. It is often used to compare the predictive performance of different models and to evaluate the improvement of the models.

3.1.4.2. MPSNR. The peak signal-to-noise ratio (often abbreviated PSNR) indicates the ratio of the maximum possible power of a signal to the destructive noise power that affects its representation accuracy. For hyperspectral images, it needs to calculate PSNR for different bands, respectively, and then take the average value, this index is called MPSNR. For a single image, the formula for calculating PSNR is:

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX^2}{MSE} \right) \quad (14)$$

MAX is the maximum pixel value possible for an image. MSE is the Mean Squared Error, defined as:

$$MSE = \frac{1}{m \cdot n} \sum_{i=1}^m \sum_{j=1}^n (A(i,j) - B(i,j))^2 \quad (15)$$

where, $A(i,j)$ and $B(i,j)$ are the pixel values of the reference image and the reconstructed image at the position (i,j) , respectively, and m and n are the height and width of the image.

For multiple images, the MPSNR is the average of the PSNR of all images. There are N images, the reference image is $\{A_1, A_2, \dots, A_N\}$, and the reconstructed image is $\{B_1, B_2, \dots, B_N\}$, then the formula for MPSNR is:

$$MPSNR = \frac{1}{N} \sum_{i=1}^N PSNR(A_i, B_i) \quad (16)$$

The larger the MPSNR value, the better the quality of the image, generally speaking: (1) higher than 40 dB: indicates that the image quality is excellent (that is, very close to the original image); (2) 30–40 dB: usually means that the image quality is good (that is, the distortion can be detected but acceptable); (3) 20–30 dB: indicates poor image quality; (4) Below 20 dB: the image quality is unacceptable.

3.1.4.3. SAM. Spectral angle mapping (often abbreviated SAM) measures the similarity between the spectra by calculating the angle between the two vectors. Specifically, SAM is calculated by the following formula:

$$SAM(A, B) = \frac{1}{M} \sum_{j=1}^M \arccos \frac{A_j \cdot B_j}{\|A_j\|_2 \|B_j\|_2} \quad (17)$$

where M represents the number of image elements of the spectrum, \cdot denotes the inner product of two vectors. The smaller SAM indicates the better fusion effect.

It is used to evaluate the spectral distortion index. The smaller the SAM value is, the smaller the spectral distortion degree is and the better the fusion result is.

3.1.4.4. SSIM. SSIM is the abbreviation of Structural Similarity, indicating structural similarity in the range of $[-1, 1]$. Specifically, SSIM is calculated by the following formula:

$$SSIM(A, B) = \frac{1}{E} \sum_{i=1}^E \frac{(2\mu A^i \mu B^i)(2\sigma A^i B^i + C_2)}{(\mu^2 A^i + \mu^2 B^i + C_1)(\sigma^2 A^i + \sigma^2 B^i + C_2)} \quad (18)$$

where E denotes the number of image bands, μA^i and μB^i represent the mean values of A and B , $\sigma^2 A^i$ and $\sigma^2 B^i$ represent the variances of A and B , respectively, $\sigma A^i B^i$ represents the covariance between A and B , C_1 and C_2 are two constants.

The closer it is to 1, the higher the similarity and the better the fusion quality.

3.1.4.5. ERGAS. ERGAS is defined as the average of the relative error between the actual observed value and the predicted value of the model. Specifically, the ERGAS relative global dimensionless error is calculated by the following formula:

$$ERGAS(A, B) = 100 \sqrt{\frac{1}{S} \sum_{i=1}^S \left(\frac{RMSE(A_i, B_i)}{\mu(B_i)} \right)^2} \quad (19)$$

where N is the number of observed values, A is the observed value, and B is the predicted value of the model. This formula averages the relative error between the actual observed value and the predicted value of the model to obtain a global error measure.

ERGAS relative global dimensionless error is often used to compare the predictive performance of different models and to evaluate the improvement of the model. The closer the value of this index is to 0, the higher the prediction accuracy of the model.

3.2. Experimental results of Simulated data

Figure 5 shows the fusion results of different fusion methods on the Simulated Dataset-1. From the figure, it can be known that the results of UFSURN-LS and UDCNN have obvious serious spectral distortion. The results of EDCSTFN, UFSURN-SL and UMC2FF retain good spatial and spectral details on the whole, but the spectral distortion is still present in the local area. The result of UFSURN has an excellent performance in spectral preservation and spatial detail compared with previous models. This proves the effectiveness and reliability of the proposed model compared with previous models.

As shown in Table 2, the proposed model outperforms all baseline methods on MPSNR, SSIM and ERGAS metrics, is in the first echelon on RMSE metric, and slightly weaker than UFSURN-LS on SAM metric. This implies a risk of spectral detail loss may exist in deeper networks. But combining the results of all metrics, although UFSURN is slightly worse in spectral angular fidelity, it is more competitive in spatial feature extraction and overall performance. This shows that the proposed UFSURN has applicability and effective compared with the previous methods. Among them, the UFSURN-LS and the UFSURN-SL are the models by performing two additional modifications of the proposed method, namely ablation experiments. As can be seen from the experimental performance, the

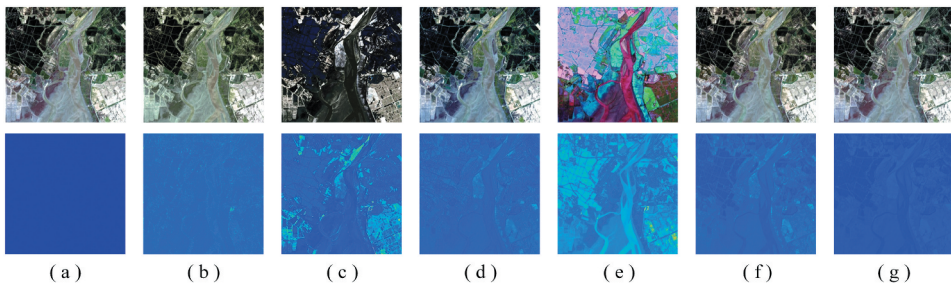


Figure 5. Fusion results of four methods on the Simulated Dataset-1. (a) HSI observed on 13 May 2022, (b) EDCSTFN, (c) UDCNN, (d) UMC2FF, (e) UFSURN-LS, (f) UFSURN-SL, (g) UFSURN.

Table 2. Quantitative assessment of comparison methods on the Simulated Dataset-1.

Methods	RMSE	MPSNR	SAM	SSIM	ERGAS
EDCSTFN	0.0387	33.0105	24.0744	0.7991	24.9165
UDCNN	0.0702	14.8823	13.9076	0.7147	15.9076
UMC2FF	0.0106	36.7637	5.5368	0.9774	3.4859
UFSURN-LS	0.0987	13.2037	23.8092	0.2558	21.0591
UFSURN-SL	0.0103	36.6757	4.7400	0.9612	3.5248
UFSURN	0.0104	39.3094	4.8216	0.9797	3.3908

model with the change of nonlinear variation and the addition of the SUREsNet has a significant good effect on most metrics.

Figure 6 shows the fusion results of different fusion methods on the Simulated Dataset-2. From the figure, it can be known the result of UFSURN-LS has the most severe global spectral distortion. The results of EDCSTFN and UFSURN-SL have obvious colour differences compared with reference image, and the visualization results are poor. The result of UMC2FF retains good spatial and spectral details on the whole, but the spectral distortion is still present in the local area. The result of UDCNN is obviously different from the reference images, and there are some spectral distortions in some parts. UFSURN demonstrates superior performance in both spectral preservation and spatial detail, highlighting its effectiveness and reliability over previous models. This proves the effectiveness and reliability of the proposed model compared with previous models.

Table 3 shows the quantitative fusion performance of different fusion methods on Simulated Dataset-2. UFSURN achieves the best value of all metrics. EDCSTFN has the worst SAM and the worst ERGAS. UMC2FF has better MPSNR and SSIM in all benchmark methods. UFSURN-LS has the worst SSIM. UFSURN with $SAM < 5$, $ERGAS < 2$, $MPSNR > 39$, and $RMSE < 0.1$ is significantly superior to these benchmark methods. The proposed UFSURN maintained optimal performance on both Simulated Dataset-1 ($MPSNR = 39.3094$, $ERGAS = 3.3908$) and Simulated Dataset-2 ($MPSNR = 39.2188$, $ERGAS = 1.6287$), with small fluctuations in quantitative indicators.

This proves the model's stability in different phenological periods and feature scenarios under the ZY1-02D sensor, further verifying the representativeness of the simulated dataset.

3.3. Experimental results of real data

Due to the lack of ground truth for HRHSI at the study site, this paper uses the HRMSI as a reference for spatial detail evaluation. At the same time, this paper plots some of the

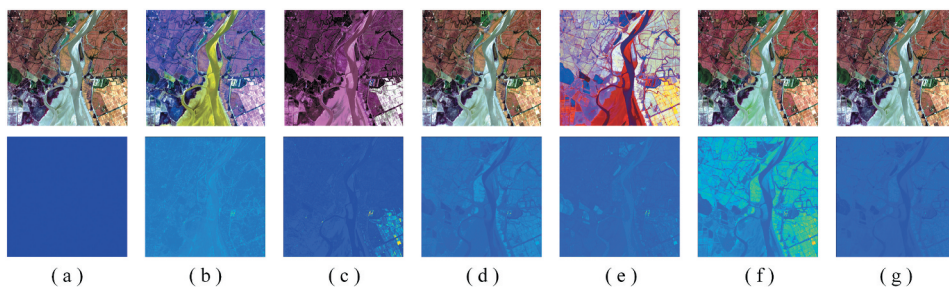


Figure 6. Fusion results of four methods on the Simulated Dataset-2. (a) HSI observed on 22 October 2022, (b) EDCSTFN, (c) UDCNN, (d) UMC2FF, (e) UFSURN-LS, (f) UFSURN-SL, (g) UFSURN.

Table 3. Quantitative assessment of comparison methods on the Simulated Dataset-2.

Methods	RMSE	MPSNR	SAM	SSIM	ERGAS
EDCSTFN	0.0561	33.3832	30.7152	0.7688	44.0813
UDCNN	0.1134	14.4678	15.5745	0.6537	19.2627
UMC2FF	0.1095	39.0241	4.9554	0.9810	1.6223
UFSURN-LS	0.0602	15.8420	28.5846	0.3751	20.6338
UFSURN-SL	0.0120	37.1778	5.4490	0.9725	3.2055
UFSURN	0.0105	39.2188	4.7939	0.9811	1.6287

spectral curves of some typical features and corresponding SAM (spectral angle mapping) visualization results which will demonstrate the excellent performance of our fusion result in spectral diagnostics.

Figure 7 illustrates the fusion results of different methods on the real dataset-1. It shows that all methods successfully capture significant spatial details. However, there are obvious differences in the spectra of the fused images. The EDCSTFN method does not perform well in predicting changes in spectral information and exhibits severe spectral distortion. Although EDCSTFN, UFSURN-SL, UDCNN and UMC2FF perform well in the prediction of the overall spectral information, the spectra of local areas, such as paddy fields, Suaeda salsa and other ground objects still show certain spectral distortion. Compared with the previous models, the proposed UFSURN performs better in the prediction of spatial details and spectral information.

To visually evaluate the performance of the proposed model, this paper compares the spectral curves of typical ground objects in dataset. Additionally, to demonstrate the effectiveness of spectral diagnosis, the spectral curves of ground objects from LRHSI on 19 March 2022, and 13 May 2022, are also included. Figure 8 displays the spectral curves featuring typical characteristics of *Phragmites australis* and *Suaeda salsa*. The results indicate that the spectral curves reconstructed by UFSURN-LS consistently hover around 0. The spectral curves reconstructed by the EDCSTFN method exhibit significant differences from the actual ground object spectral curves. Although the spectral curves reconstructed by the UFSURN-SL, UDCNN, and UMC2FF

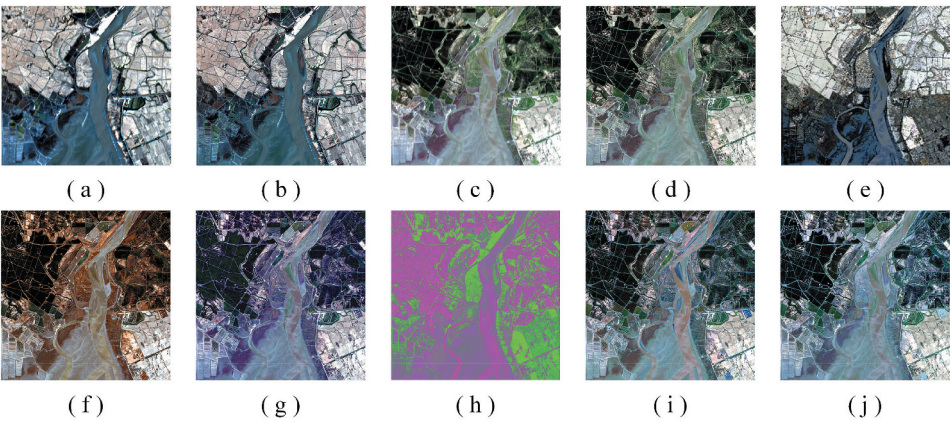


Figure 7. Fusion results of four methods on the real dataset-1. (a) HSI observed on 19 March 2022, (b) MSI observed on 19 March 2022, (c) HSI observed on 13 May 2022, (d) MSI observed on 13 May 2022, (e) EDCSTFN, (f) UDCNN, (g) UMC2FF, (h) UFSURN-LS, (i) UFSURN-SL, (j) UFSURN.

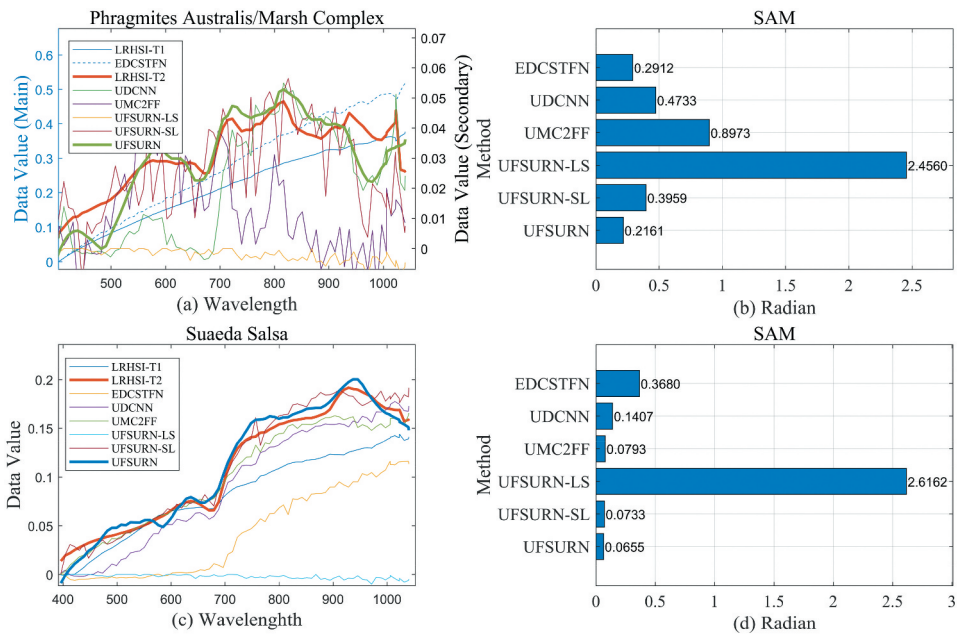


Figure 8. The spectral curves of the typical features.

methods are close to the actual spectral curves, they still show considerable spectral jitter. The spectral curves reconstructed by the proposed UFSURN method closely resemble the actual values, exhibit no significant spectral jitter, and differ markedly from the data on 6 September 2022. SAM quantifies the spectral distortion by calculating the angle between the spectral vectors, where a smaller value indicates a higher spectral similarity. Thus, this paper calculated the SAM values of the fusion results and the real spectra for further analysis and comparison. From the SAM bar chart on the right of Figure 8, it can be seen that the proposed method has the smallest SAM value in the spectral reconstruction of two typical ground objects. The above shows that the proposed method's superior performance in spectral reconstruction and diagnosis.

Figure 9 illustrates the fusion results of different methods on Real Dataset-2. It shows that all methods preserve the spatial details significantly. However, there are obvious differences in the spectra of the fused images. UFSURN-LS, UFSURN-SL and UDCNN have poor overall visualizations. The EDCSTFN method has a poor reconstruction effect in vegetated areas. The UMC2FF method has a poor visualization effect in the aquaculture ponds. Among all the methods, the UFSURN method achieves the best reconstruction results.

Similarly, in Real Dataset-2, this paper selects two typical features: *Phragmites australis* and intertidal muds. Figure 10 shows that the spectral curves reconstructed by UFSURN-LS, UFSURN-SL, UDCNN and UMC2FF are significantly different from those of the actual ground objects. The spectral curve reconstructed by EDCSTFN is close to the spectral curve at T_1 , but has strong spectral jitter. In contrast, the spectral curves reconstructed by the UFSURN method are very close to the actual values without significant spectral jitter.

Besides, the bar chart of SAM visualization on the right shows that UFSURN has the lowest SAM in the spectral reconstruction of two typical ground objects. The above shows that the proposed method has richer spectral details and more accurate reconstruction results compared to previous methods.

4. Conclusion

Recent advances in STSF methods have made significant contributions to the field. By integrating spatial-spectral fusion and spectral-temporal fusion into a unified progressive end-to-end framework, PSSTFN (X. Chen et al. 2023) effectively captures nonlinear spatial-temporal-spectral relationships and achieves excellent performance in scenarios with sufficient high-quality labelled data. However, these methods still face limitations in practical applications for domestic low-temporal-resolution hyperspectral data (e.g. ZY1-02D), where the scarcity of labelled HRHSI and the mismatch between data temporal attributes and method assumptions restrict their applicability. For DCSTFN: This method is mainly designed for STSF of MODIS (high temporal resolution) and Landsat (medium spatial resolution) data, assuming that hyperspectral data has higher temporal resolution than multispectral data. In contrast, our study focuses on ZY1-02D (low temporal resolution HSI) and ZY1-02D (high temporal resolution MSI) data, which has the opposite temporal resolution attribute of input data. In order to solve the problem that the existing STSF methods are not suitable for domestic hyperspectral satellite data, and the nonlinear relationship between time, space and spectrum, based on the lack of hyperspectral training data, this paper proposes an unsupervised STSF method for hyperspectral images (UFSURN). The main contributions are as follows: The STSF method proposed in this paper, which completely relies on the LRHSI and HRMSI at time 1 and HRMSI at time 2, and does not require the participation of other additional data, which is in line with the realistic situation of lacking enough hyperspectral training data. At the same time, the spectral upsampling network uses a residual network with spectral channel adaptation. This design can effectively avoid the degradation of output

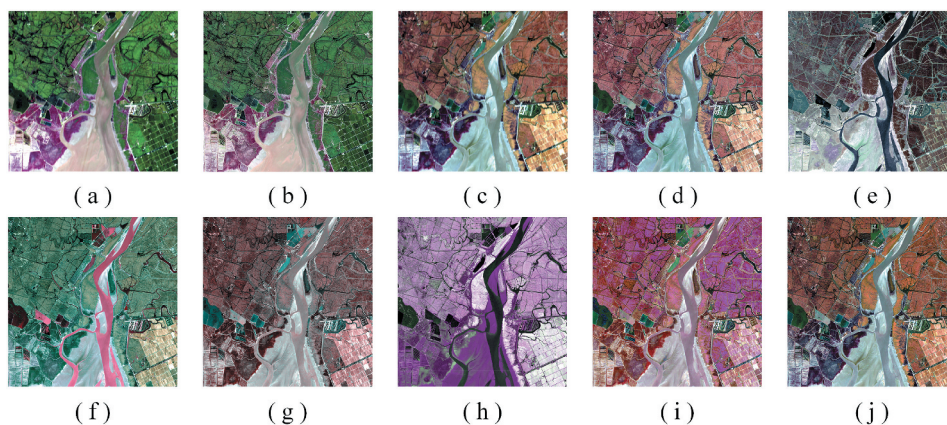


Figure 9. Fusion results of four methods on the real dataset-2. (a) HSI observed on 06 September 2022, (b) MSI observed on 06 September 2022, (c) HSI observed on 22 October 2022, (d) MSI observed on 22 October 2022, (e) EDCSTFN, (f) UDCNN, (g) UMC2FF, (h) UFSURN-LS, (i) UFSURN-SL, (j) UFSURN.

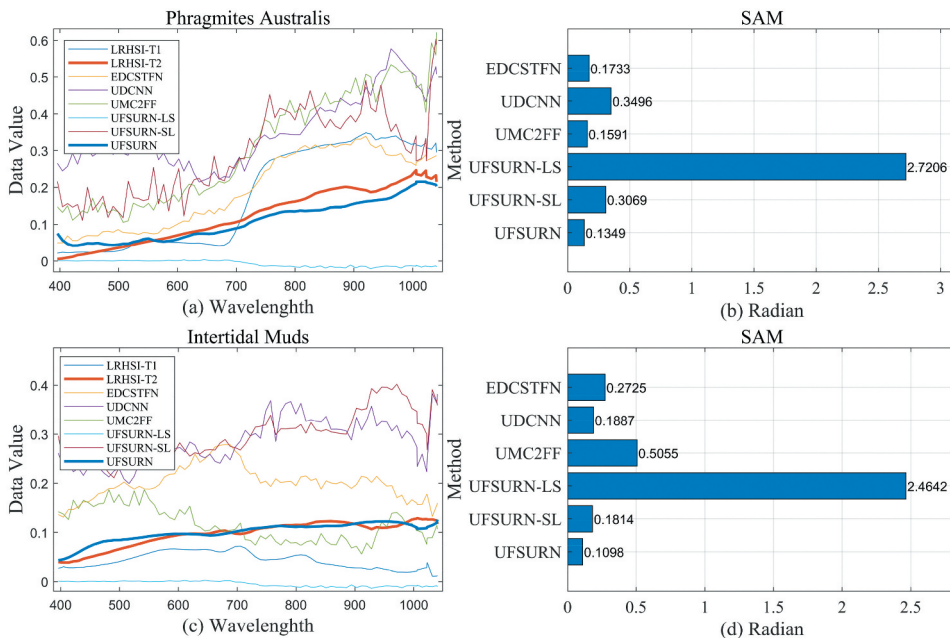


Figure 10. The spectral curves of the typical features.

results and realize spectral adaptive upsampling while deepening the depth of the network. The spatial-spectral downsampling network uses a global shared convolutional neural network to accurately describe the nonlinear constraint relationship of the spatial-spectral degradation model. The fusion results have been verified by simulated and real data sets based on MSI and HSI data of ZY1-02D. The experimental results show that the proposed method has richer spatial and spectral details and more accurate reconstruction results.

Although the proposed model performs well in several test scenarios, its performance is still constrained by certain assumptions and limitations. Specifically, the performance of this model under extreme conditions, such as large-scale deformations, drastic illumination changes, etc., can be significantly affected, as such models usually rely on the uniformity and stability of the data to some extent. In order to improve the generalization ability and reliability of the model in practice, future research can consider how to enhance the robustness of the model under non-ideal conditions. Future research can focus on the following areas: Firstly, the robustness and flexibility of the model are enhanced by expanding the training data to adapt to more complex environmental conditions. Secondly, explore the improvement of model performance by new technologies such as deep learning. Finally, interdisciplinary approaches can be considered to open up new research areas by combining remote sensing techniques with other data sources such as climate models, etc.

Acknowledgements

The authors would like to thank Dr Yao Liu from the Landsat Remote Sensing Application Center of the Ministry of Natural Resources of China for generously providing the ZY1-02D data.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported in part by National Natural Science Foundation of China under Grant [42471380], Grant [42201362], and Grant [42371353], in part by Joint Open Fund of the Research Platforms of School of Computer Science, China University of Geosciences, Wuhan, under Grant [PTLH2024-B-05], in part by Shandong Provincial Natural Science Foundation of China under Grant [ZR2024QD042], and in part by the Fundamental Research Funds for Central University Basic Research Fund of China under Grant [3132025264].

ORCID

Haoyang Yu  <http://orcid.org/0000-0002-4026-7450>

Data availability statement

The LN-02T dataset produced by the authors used in this article is publicly available in <https://github.com/Welcome-to-LISA/LN02T>.

References

- Chen, B., B. Huang, and B. Xu. 2015. "Comparison of Spatiotemporal Fusion Models: A Review." *Remote Sensing* 7 (2): 1798–1835. <https://doi.org/10.3390/rs70201798>.
- Chen, X., X. Meng, F. Shao, and W. Sun. 2023. "PSSTFN: A Progressive Spatial–Temporal–Spectral Fusion Network for Remote Sensing Images." *IEEE Transactions on Geoscience & Remote Sensing* 61:1–12. <https://doi.org/10.1109/TGRS.2023.3329531>.
- Dian, R., S. Li, and L. Fang. 2016. "Non-Local Sparse Representation for Hyperspectral Image Super-Resolution." Paper presented at the 2016 IEEE International Conference on Image Processing (ICIP) Phoenix, AZ. 25–28 September. 2016.
- Dong, W., F. Fu, G. Shi, X. Cao, J. Wu, G. Li, and X. Li. 2016. "Hyperspectral Image Super-Resolution via Non-Negative Structured Sparse Representation." *IEEE Transactions on Image Processing* 25 (5): 2337–2352. <https://doi.org/10.1109/TIP.2016.2542360>.
- Feng, G., J. Masek, M. Schwaller, and F. Hall. 2006. "On the Blending of the Landsat and MODIS Surface Reflectance: Predicting Daily Landsat Surface Reflectance." *IEEE Transactions on Geoscience & Remote Sensing* 44 (8): 2207–2218. <https://doi.org/10.1109/TGRS.2006.872081>.
- Hou, J., X. Liu, C. Wu, X. Cong, C. Huang, L.-J. Deng, and J. Wei You. 2025. "Bidomain Uncertainty Gated Recursive Network for Pan-sharpening." *Information Fusion* 118:102938. <https://doi.org/10.1016/j.inffus.2025.102938>.
- Huang, B., Z. Hankui, S. Huihui, W. Juan, and C. Song. 2013. "Unified Fusion of Remote-Sensing Imagery: Generating Simultaneously High-Resolution Synthetic Spatial–Temporal–Spectral Earth Observations." *Remote Sensing Letters* 4 (6): 561–569. <https://doi.org/10.1080/2150704X.2013.769283>.
- Huang, et al. 2013a. And it refers to the following reference article. Thank you for your help, and please help to modify it. Huang, B., J. Wang, H. Song, D. Fu, and K. Wong. 2013. "Generating High Spatiotemporal Resolution Land Surface Temperature for Urban Heat Island Monitoring." *IEEE Geoscience & Remote Sensing Letters* 10 (5): 1011–1015. <https://doi.org/10.1109/LGRS.2012.2227930>.

- Huang, B., J. Wang, H. Song, D. Fu, and K. Wong. 2013b. "Generating High Spatiotemporal Resolution Land Surface Temperature for Urban Heat Island Monitoring." *IEEE Geoscience & Remote Sensing Letters* 10 (5): 1011–1015. <https://doi.org/10.1109/LGRS.2012.2227930>.
- Jia, J., H. Yu, C. Wang, K. Zheng, J. Li, and J. Hu. 2025. "Spectral-Spatial Collaborative Pretraining Framework with Multiconstraint Cooperation for Hyperspectral–Multispectral Image Fusion." *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 18:11610–11622. <https://doi.org/10.1109/JSTARS.2025.3562278>.
- Jiang, M., H. Shen, and J. Li. 2022. "Deep-Learning-Based Spatio-Temporal-Spectral Integrated Fusion of Heterogeneous Remote Sensing Images." *IEEE Transactions on Geoscience & Remote Sensing* 60:1–15. <https://doi.org/10.1109/TGRS.2022.3188998>.
- Li, A., Y. Bo, Y. Zhu, P. Guo, J. Bi, and Y. He. 2013. "Blending Multi-Resolution Satellite Sea Surface Temperature (SST) Products Using Bayesian Maximum Entropy Method." *Remote Sensing of Environment* 135:52–63. <https://doi.org/10.1016/j.rse.2013.03.021>.
- Li, D., Y. Li, W. Yang, Y. Ge, Q. Han, L. Ma, Y. Chen, and X. Li. 2018. "An Enhanced Single-Pair Learning-Based Reflectance Fusion Algorithm with Spatiotemporally Extended Training Samples." *Remote Sensing* 10 (8). <https://doi.org/10.3390/rs10081207>.
- Li, J., D. Hong, L. Gao, J. Yao, K. Zheng, B. Zhang, and J. Chanussot. 2022. "Deep Learning in Multimodal Remote Sensing Data Fusion: A Comprehensive Review." *International Journal of Applied Earth Observation and Geoinformation* 112:102926. <https://doi.org/10.1016/j.jag.2022.102926>.
- Li, J., Y. Li, L. He, J. Chen, and A. Plaza. 2020. "Spatio-Temporal Fusion for Remote Sensing Data: An Overview and New Benchmark." *Science China Information Sciences* 63 (4): 140301. <https://doi.org/10.1007/s11432-019-2785-y>.
- Li, J., K. Zheng, W. Liu, Z. Li, H. Yu, and L. Ni. 2023. "Model-Guided Coarse-to-Fine Fusion Network for Unsupervised Hyperspectral Image Super-Resolution." *IEEE Geoscience & Remote Sensing Letters* 20:1–5. <https://doi.org/10.1109/LGRS.2023.3309854>.
- Li, Y., J. Li, L. He, J. Chen, and A. Plaza. 2020. "A New Sensor Bias-Driven Spatio-Temporal Fusion Model Based on Convolutional Neural Networks." *Science China Information Sciences* 63 (4): 140302. <https://doi.org/10.1007/s11432-019-2805-y>.
- Maselli, F., and F. Rembold. 2002. "Integration of LAC and GAC NDVI Data to Improve Vegetation Monitoring in Semi-Arid Environments." *International Journal of Remote Sensing* 23 (12): 2475–2488. <https://doi.org/10.1080/01431160110104755>.
- Qu, Y., H. Qi, and C. Kwan. 2018. "Unsupervised Sparse Dirichlet-Net for Hyperspectral Image Super-Resolution." Paper presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18–23, 2018 Salt Lake City, UT.
- Rahmani, S., M. Strait, D. Merkurjev, M. Moeller, and T. Wittman. 2010. "An Adaptive IHS Pan-Sharpening Method." *IEEE Geoscience & Remote Sensing Letters* 7 (4): 746–750. <https://doi.org/10.1109/LGRS.2010.2046715>.
- Richard, B. G., J. Amin, and K. Menas. 2001. "Wavelet-Based Hyperspectral and Multispectral Image Fusion." Paper presented at the Proc.SPIE Orlando, FL.
- Simões, M., J. Bioucas-Dias, L. B. Almeida, and J. Chanussot. 2015. "A Convex Formulation for Hyperspectral Image Superresolution via Subspace-Based Regularization." *IEEE Transactions on Geoscience & Remote Sensing* 53 (6): 3373–3388. <https://doi.org/10.1109/TGRS.2014.2375320>.
- Sun, W., K. Ren, X. Meng, G. Yang, J. Peng, and J. Li. 2023. "Unsupervised 3-D Tensor Subspace Decomposition Network for Spatial–Temporal–Spectral Fusion of Hyperspectral and Multispectral Images." *IEEE Transactions on Geoscience & Remote Sensing* 61:1–17. <https://doi.org/10.1109/TGRS.2023.3324028>.
- Sun, W., G. Yang, C. Chen, M. Chang, K. Huang, X. Meng, and L. Liu. 2020. "Development Status and Literature Analysis of China's Earth Observation Remote Sensing Satellites." *National Remote Sensing Bulletin* 24 (5): 479–510. <https://doi.org/10.11834/jrs.20209464>.
- Sun, Y., Z. Hua, and W. Shi. 2019. "A Spatio-Temporal Fusion Method for Remote Sensing Data Using a Linear Injection Model and Local Neighbourhood Information." *International Journal of Remote Sensing* 40 (8): 2965–2985. <https://doi.org/10.1080/01431161.2018.1538585>.

- Sylla, D., A. Minghelli-Roman, P. Blanc, A. Mangin, and O. Hembise Fanton d'Andon. 2014. "Fusion of Multispectral Images by Extension of the Pan-Sharpening ARSIS Method." *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 7 (5): 1781–1791. <https://doi.org/10.1109/JSTARS.2013.2271911>.
- Tan, Z., L. Di, M. Zhang, L. Guo, and M. Gao. 2019. "An Enhanced Deep Convolutional Model for Spatiotemporal Image Fusion." *Remote Sensing* 11 (24). <https://doi.org/10.3390/rs11242898>.
- Tan, Z., P. Yue, L. Di, and J. Tang. 2018. "Deriving High Spatiotemporal Remote Sensing Images Using Deep Convolutional Network." *Remote Sensing* 10 (7). <https://doi.org/10.3390/rs10071066>.
- Wang, Q., and P. M. Atkinson. 2018. "Spatio-Temporal Fusion for Daily Sentinel-2 Images." *Remote Sensing of Environment* 204:31–42. <https://doi.org/10.1016/j.rse.2017.10.046>.
- Wang, Q., Y. Zhang, A. O. Onojeghuo, X. Zhu, and P. M. Atkinson. 2017. "Enhancing Spatio-Temporal Fusion of MODIS and Landsat Data by Incorporating 250 m MODIS Data." *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 10 (9): 4116–4123. <https://doi.org/10.1109/JSTARS.2017.2701643>.
- Wei, J., L. Wang, P. Liu, X. Chen, W. Li, and A. Y. Zomaya. 2017. "Spatiotemporal Fusion of MODIS and Landsat-7 Reflectance Images via Compressed Sensing." *IEEE Transactions on Geoscience & Remote Sensing* 55 (12): 7126–7139. <https://doi.org/10.1109/TGRS.2017.2742529>.
- Wu, M., W. Huang, Z. Niu, and C. Wang. 2015. "Generating Daily Synthetic Landsat Imagery by Combining Landsat and MODIS Data." *Sensors* 15 (9): 24002–24025. <https://doi.org/10.3390/s150924002>.
- Yan, J., K. Zhang, Q. Sun, C. Ge, W. Wan, J. Sun, and H. Zhang. 2025. "Spatial-Spectral Unfolding Network with Mutual Guidance for Multispectral and Hyperspectral Image Fusion." *Pattern Recognition* 161:111277. <https://doi.org/10.1016/j.patcog.2024.111277>.
- Yang, H., H. Yu, K. Zheng, J. Hu, T. Tao, and Q. Zhang. 2023. "Hyperspectral Image Classification Based on Interactive Transformer and CNN with Multilevel Feature Fusion Network." *IEEE Geoscience & Remote Sensing Letters* 20:1–5. <https://doi.org/10.1109/LGRS.2023.3303008>.
- Yao, J., D. Hong, J. Chanussot, D. Meng, X. Zhu, and Z. Xu. 2020. "Cross-Attention in Coupled Unmixing Nets for Unsupervised Hyperspectral Super-Resolution." Paper Presented at the Computer Vision – ECCV 2020, Cham, 2020.
- Yu, H., Z. Ling, K. Zheng, L. Gao, J. Li, and J. Chanussot. 2024. "Unsupervised Hyperspectral and Multispectral Image Fusion with Deep Spectral-Spatial Collaborative Constraint." *IEEE Transactions on Geoscience & Remote Sensing* 62:1–14. <https://doi.org/10.1109/TGRS.2024.3472226>.
- Yu, H., Z. Ling, K. Zheng, J. Li, S. Liang, and L. Gao. 2023. "Unsupervised Dynamic Convolutional Neural Network Model for Hyperspectral and Multispectral Image Fusion." Paper presented at the IGARSS 2023–2023 IEEE International Geoscience and Remote Sensing Symposium, July 16–21, 2023 Pasadena, CA.
- Yu, H., H. Yang, L. Gao, J. Hu, A. Plaza, and B. Zhang. 2024. "Hyperspectral Image Change Detection Based on Gated Spectral–Spatial–Temporal Attention Network with Spectral Similarity Filtering." *IEEE Transactions on Geoscience & Remote Sensing* 62:1–13. <https://doi.org/10.1109/TGRS.2024.3373820>.
- Zhang, L., D. Fu, X. Sun, H. Chen, and X. She. 2016. "A Spatial-Temporal-Spectral Blending Model Using Satellite Images." *IOP Conference Series Earth and Environmental Science* 34 (1): 012042. <https://doi.org/10.1088/1755-1315/34/1/012042>.
- Zhang, Y., S. De Backer, and P. Scheunders. 2009. "Noise-Resistant Wavelet-Based Bayesian Fusion of Multispectral and Hyperspectral Images." *IEEE Transactions on Geoscience & Remote Sensing* 47 (11): 3834–3843. <https://doi.org/10.1109/TGRS.2009.2017737>.
- Zhang, Y., and M. He. 2007. "Multi-Spectral and Hyperspectral Image Fusion Using 3-D Wavelet Transform." *Journal of Electronics (China)* 24 (2): 218–224. <https://doi.org/10.1007/s11767-005-0232-5>.
- Zhao, Y., and B. Huang. 2017. "Integrating MODIS and MTSAT-2 to Generate High Spatial-Temporal-Spectral Resolution Imagery for Real-Time Air Quality Monitoring." Paper presented at the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) Fort Worth, Texas. July 23–28, 2017.

- Zheng, K., L. Gao, W. Liao, D. Hong, B. Zhang, X. Cui, and J. Chanussot. 2021. "Coupled Convolutional Neural Network with Adaptive Response Function Learning for Unsupervised Hyperspectral Super Resolution." *IEEE Transactions on Geoscience & Remote Sensing* 59 (3): 2487–2502. <https://doi.org/10.1109/TGRS.2020.3006534>.
- Zheng, K., A. Khader, and L. Xiao. 2022. "An Unsupervised Hyperspectral Image Fusion Method Based on Spectral Unmixing and Deep Learning." Paper presented at the IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, July 17–22, 2022 Kuala Lumpur, Malaysia.
- Zhou, J., W. Sun, X. Meng, G. Yang, K. Ren, and J. Peng. 2022. "Generalized Linear Spectral Mixing Model for Spatial–Temporal–Spectral Fusion." *IEEE Transactions on Geoscience & Remote Sensing* 60:1–16. <https://doi.org/10.1109/TGRS.2022.3188501>.
- Zhu, C., D. Shangqi, L. Jiaxin, Z. Ying, G. Liwei, G. Liangbo, T. Na, C. Shengbo, and Q. Wu. 2023. "Hyperspectral and Multispectral Remote Sensing Image Fusion Using SWINGAN with Joint Adaptive Spatial-Spectral Gradient Loss Function." *International Journal of Digital Earth* 16 (1): 3580–3600. <https://doi.org/10.1080/17538947.2023.2253206>.
- Zhu, C., T. Zhang, Q. Wu, Y. Li, and Q. Zhong. 2024. "An Implicit Transformer-Based Fusion Method for Hyperspectral and Multispectral Remote Sensing Image." *International Journal of Applied Earth Observation and Geoinformation* 131:103955. <https://doi.org/10.1016/j.jag.2024.103955>.
- Zurita-Milla, R., J. G. P. W. Clevers, and M. E. Schaepman. 2008. "Unmixing-Based Landsat TM and MERIS FR Data Fusion." *IEEE Geoscience & Remote Sensing Letters* 5 (3): 453–457. <https://doi.org/10.1109/LGRS.2008.919685>.