Spatial–Spectral Cross Fusion Attention based Hyperspectral Image Super-Resolution for Land Resource Auditing

Jinjin Zhang[1], Yu Wang[1], Ranchen Dai[1], Tianming Zhan[1] and Xiaobing Yu[1*]
[1] School of Computer Science and School of Intelligence Audit, Nanjing Audit University, Nanjing, 211815 China
E-mail: zhangjj1981@sohu.com
*Corresponding author

Keywords: hyperspectral image, super-resolution, transformer, cross-attention fusion

Received: November 20, 2024

Hyperspectral imaging, celebrated for its detailed spectral information, finds broad application in various fields. However, the limitations inherent to optical systems often impede the direct acquisition of high-resolution hyperspectral images. Hence, achieving these images has become a key focus in the research community. The process of single hyperspectral image super-resolution (HSI-SR) aims to upscale low-resolution images to a higher resolution. With the evolution of deep learning, the incorporation of Convolutional Neural Networks (CNNs) into super-resolution methods has shown considerable promise. Yet, the challenge lies in the thorough extraction of both spatial and spectral data, especially in the context of remote sensing, which can limit the model's ability to learn effectively. Additionally, Transformer-based techniques often struggle to capture the intricate relationships between spatial and spectral features, which can hinder the effectiveness of image reconstruction. To overcome these challenges, this paper presents a novel HSI-SR approach: Spatial–Spectral Cross Fusion Attention based Hyperspectral Image Super-Resolution for Land Resource Auditing, which synergizes the strengths of CNNs and the Transformer architecture. During the learning of spatial features, the method alternates between window self-attention and zero-padding window self-attention, allowing for a more comprehensive focus on feature information and the integration of different windows to achieve long-range insights. Furthermore, the cross-attention feature fusion module designed for this approach is adept at merging spatial and spectral features, thus enhancing the model's ability to learn from both types of information. The approach effectively enhances spatial-spectral integration, improving reconstruction quality. Extensive experimental assessments have demonstrated the proposed method's superiority over current industry benchmarks. PSNR improvements 0.08 over baseline in Cave.

Povzetek: Predstavljen je nov pristop za izboljšanje prostorsko-spektralne ločljivosti hiperspektralnih slik z uporabo križno-pozornostnega združevanja, kar izboljša kakovost rekonstruiranih slik za revizijo zemljiških virov.

1 Introduction

Hyperspectral imaging, characterized by its continuous narrow-band data and high spectral resolution, offers an abundance of spectral information that can discern the subtlest of spectral features. These capabilities have been widely utilized across various domains, including construction audits [1], Hyperspectral imaging technology, with its continuous narrowband spectral data and high spectral resolution, reveals unprecedented spectral information details and can accurately capture the slightest spectral differences [2], [3], [4]. The unique advantages of this technology have been widely recognized and applied in multiple fields, especially in the field of "land resource auditing" [5], [6], [7], where it plays an irreplaceable role [8], [9]. On the other hand, imaging systems designed specifically for high spatial resolution often have smaller IFOVs, which in turn require wider spectral channels to collect sufficient light energy[10], [11], [12]. However, in the realm of remote sensing imaging systems, there is often a necessary compromise to be struck between achieving high spatial resolution and capturing detailed spectral information. The narrow spectral bandwidth inherent to hyperspectral imaging systems necessitates a large instantaneous field of view (IFOV) to gather sufficient light quanta, thereby ensuring a satisfactory signal-to-noise ratio. Conversely, systems designed for high spatial resolution feature a smaller IFOV, which in turn demands a broader spectral channel. Consequently, current remote sensing imaging systems frequently fall short of delivering both high spatial and spectral resolutions simultaneously. This limitation restricts the broader application of hyperspectral images across various domains. For example, low-resolution images can impact the audit process in land resource audits. Consequently, the development of methods to obtain hyperspectral images with enhanced spatial resolution has emerged as a pivotal research direction.

Image Super-Resolution, a crucial technique in image enhancement [13], [14], [15], [16], enables the reconstruction of high-resolution images. Classified by the

quantity of input images, this technique bifurcates into fusion-based [17], [18], [19] and single hyperspectral image super-resolution approaches. The model optimization method describes the relationship between high-resolution multispectral images (HR MSI) and low-resolution hyperspectral images by constructing a degradation model, in order to more accurately reflect the complex degradation process in the real world. This method often requires a combination of appropriate prior information and constraints to derive the target image through optimization algorithms [23]. The fusion method based on deep learning fully utilizes the spatial and spectral correlations between LR HSI and HR MSI to further improve the accuracy of super-resolution reconstruction. Improving the resolution of HSI is of crucial importance in the field of land resource auditing. Land resource auditing requires precise assessment of surface cover, land use status, and soil characteristics, while hyperspectral images can provide rich spectral information that helps identify different types of vegetation, soil, and man-made objects [24]. However, due to the limitations of remote sensing imaging systems, the obtained hyperspectral images often have low resolution, making it difficult to meet the detailed information requirements of land resource audits. Therefore, improving the resolution of images through the HSI-SR method can significantly enhance the accuracy and efficiency of land resource auditing. However, these fusion-based techniques necessitate auxiliary high-resolution multispectral images, imposing certain prerequisites on the quality of the supplementary data. The assumption of a strong correlation between input images, which is often a prerequisite for most fusion methods, poses a practical challenge due to the difficulty of acquiring well-matched images, thereby constraining their real-world applicability.

In stark contrast to fusion methods, single-frame hyperspectral image super-resolution eschews the need for auxiliary information, opting to directly upscale a LR HSI to a HR HSI, thereby enhancing its practical viability in real-world scenarios. Principal techniques within this domain encompass interpolation, low-rank tensor approximation [25], sparse representation [26], and deep learning [27]. Interpolation techniques, which estimate unknown pixel values based on their neighbors, often fall short in capturing high-frequency details, leading to edge blurring. To delve deeper into the intrinsic characteristics of hyperspectral images, tensor completion-based methods have been proposed for spatial super-resolution, albeit at the cost of computational efficiency due to their formulation as complex iterative optimization problems. Deep learning-based SR methods have demonstrated remarkable efficacy, attracting substantial research interest. The objective of these techniques is to identify the complex relationships between images of low and high spatial resolution for the purpose of hyperspectral image reconstruction, with Convolutional Neural Networks and Transformer models becoming prominent approaches in this field.

The rapid evolution of deep learning has endowed CNNs with the ability to extract and learn profound image features through convolutional, pooling, and fully connected layers, thereby achieving remarkable success in image classification [28], [29], [30], object detection [31], [32], [33], and beyond. SRCNN [34] marked a seminal application of CNNs in image super-resolution, significantly improving the reconstruction of natural images over conventional techniques. Building upon this, advanced methods integrating residual learning [35] and multi-scale processing have surfaced, enhancing the capacity of the model to learn complex features.

As deep learning technology progresses, the Transformer architecture, has made significant inroads into the realm of CV. Its self-attention mechanism, pivotal for learning key features and capturing long-range dependencies, has notably improved the reconstruction quality of high-resolution hyperspectral images (HR-HSIs). However, the single-image super-resolution process often suffers from a lack of interaction between spectral and spatial information, degrading reconstruction quality.

To counter these challenges, particularly the CNN's limitation in capturing long-range dependencies and the Transformer's struggle to integrate spatial and spectral information seamlessly, this paper introduces a Spatial–Spectral Cross Fusion Transformer which fortifies reconstruction by integrating spatial and spectral features more cohesively. The objective is to enhance spatial resolution while preserving spectral fidelity. It employs a window attention mechanism with zero-padding for spatial information to foster inter-window information exchange and enhance spatial feature capture. In the spectral realm, features are extracted through convolutional operations and a dedicated spectral attention module. Additionally, the approach enhances detail by fusing intermediate outputs from both the spatial and spectral feature extraction branches, reintegrating these refined features into subsequent modules. This promotes a robust interaction between spatial and spectral domains, achieving a more effective dimensional fusion. The culmination of this process is the amalgamation of outputs from both branches, adeptly restoring the spatial and spectral resolutions of the hyperspectral image. This not only preserves the image's full spectral and spatial integrity but also significantly bolsters the performance of hyperspectral image super-resolution.

To summarize, the key contributions of this research paper are outlined below:

- We propose a novel Spatial-Spectral Cross-Fusion Attention-Based Hyperspectral Image Super-Resolution for Land Resource Auditing. This method integrates spatial and spectral information to improve image reconstruction and enhance the accuracy of land resource auditing. The proposed framework incorporates a cross-attention fusion module that promotes effective feature interaction between spatial and spectral branches, thereby enhancing the quality of super-resolution. This method addresses the critical challenge of utilizing low-resolution hyperspectral images in land resource auditing applications.

- To better capture spatial and spectral features, we design a cross-attention feature fusion module. This module fuses the outputs of the spatial and spectral feature extraction branches, enhancing feature learning and improving the final image reconstruction effect.

2 Related work

This section offers an extensive examination of the key technological milestones within the realm of HSI-SR, encompassing the trajectory of advancements in this field. Initially, we delineate the two predominant strategies: image fusion techniques and single image SR methodologies. Following that, we provide a comprehensive review of diverse deep learning-driven methods for HSI-SR.

2.1 The methods of image super-resolution

The approaches to achieving HSI SR are predominantly classified into two main categories: those that rely on the fusion of multiple images, known as fusion-based methods, and those that enhance the resolution of a single image, referred to as single-image HSI-SR methods. The former methods necessitate supplementary information to facilitate the reconstruction process. This auxiliary information is predominantly in the form of high-resolution multispectral imagery. Such methods encompass techniques grounded in matrix decomposition, Bayesian inference, tensor factorization, and deep learning algorithms. Conversely, single-image hyperspectral SR directly upscales a LR-HSI to a high-resolution counterpart, eschewing the requirement for additional auxiliary data. Given the inherent challenges associated with procuring precise auxiliary data and mitigating spectral distortion, our study concentrates on HSI-SR techniques.

2.2 Traditional approaches

Conventional methods for image enhancement techniques predominantly utilize mathematical and signal processing approaches to augment the resolution, thereby framing the super-resolution (SR) challenge for hyperspectral images (HSI) as an optimization problem. Within this framework, diverse image priors are integrated into the optimization process to attain a favorable representation of the HSI data. Such techniques encompass interpolation, low-rank tensor approximation, sparse representation, among others. Interpolation methods, which estimate unknown pixel values based on their neighbors, often struggle to recover lost high-frequency information, resulting in blurred edges. To delve into the intrinsic characteristics of hyperspectral images, novel tensor-based methods have been introduced for enhancing resolution. Nonetheless, these methods can be computationally intensive, as they are frequently cast as complex optimization problems requiring iterative solution strategies. The inherent limitations of traditional methods have catalyzed the emergence and swift advancement of deep learning approaches. Deep learning methods offer innovative perspectives and sophisticated tools, revolutionizing the landscape of image super-resolution.

2.3 Deep learning approaches

Contrary to conventional single-image super-resolution techniques, deep learning networks excel at uncovering the intrinsic features embedded within image data, thereby offering enhanced performance in the HSI-SR domain. This section delves into the application of Convolutional Neural Networks and Transformer architectures for addressing single HSI-SR tasks.

2.3.1 CNN-based approaches

The swift evolution of deep learning has led to the successful deployment of CNNs in super-resolution techniques, yielding commendable outcomes. Dong et al. [34] pioneered the application of a three-layer CNN for natural image super-resolution, introducing the SRCNN, which amalgamates CNNs with super-resolution methods to significantly bolster image reconstruction efficacy. Motivated by these findings, subsequent research has advocated for the adaptation of similar solutions to address the super-resolution challenges specific to individual hyperspectral images. Wu et al. [39] introduced the SDCNN, employing spatial constraints to facilitate the mapping, albeit with potential performance limitations for certain image types or scenes. Li et al. [40] further proposed the GDRRN, capable of directly mapping low-resolution inputs to high-resolution outputs while adeptly capturing intricate spectral-spatial dynamics, thereby enhancing super-resolution capabilities. Nonetheless, the model's heightened complexity and extensive parameterization demand considerable computational resources and time for training and deployment, presenting a risk of overfitting.

In the realm of image super-resolution, while CNNs adeptly capture spatial features, the limitations of 2D convolution hinder the preservation of spectral information essential to hyperspectral imagery. Augmenting the network depth with residual modules further bolsters the overall image recovery process. Mei et al. [41] exemplified this with the introduction of a 3D fully convolutional neural network designed to encapsulate spectral information within its architecture, thereby capturing spatial and spectral features more effectively and enhancing super-resolution accuracy. Nonetheless, the model's efficacy fluctuates with varying hyperspectral datasets and applications, which may impede its generalizability. Li et al. [40] advanced the GDRRN by integrating grouped convolution within the recurrent residual module, effectively supplanting the traditional 3D convolution. However, these methodologies struggle to transcend the inherent focus of CNNs on local features, often overlooking the long-range dependencies present within images. These constraints can significantly impede the model's capacity for feature learning, culminating in suboptimal reconstruction results.

2.3.2 Transformer-based approaches

Transformer framework has been adopted by the field of CV due to its self-attention mechanism's ability can capture long-term dependencies among features and patches, yielding enhanced performance. Specifically, Liang et al. [42] introduced the Swin Transformer for natural image recovery, which has demonstrated superior results. However, its potential to disrupt spectral correlations renders it less suitable for hyperspectral image recovery tasks. Consequently, researchers have begun integrating 3D convolution with the Transformer to concurrently learn spatial and spectral features, thus engaging with both local and global image characteristics. Liu et al. [43] proposed the Interactformer, integrating an interactive transformer with a CNN to address hyperspectral image super-resolution. Wu et al. [44] also integrated spectral attention mechanisms with three-dimensional convolutional operations to capture the characteristics within an extensive receptive field, thereby enhancing the feature extraction process in hyperspectral imaging. However, these methods focus on the extraction of spatial and spectral features, overlooking the critical role that their interaction plays in bolstering reconstruction quality during image super-resolution. In response, we introduce the Spatial–Spectral Cross Fusion Transformer for Hyperspectral Image Super-Resolution, designed to effectively mediate information exchange and to fully harness spatial and spectral information for HSI-SR.

3 Method

This section delineates the methodology of our approach. Section 3.1 outlines the architecture of the entire network. Section 3.2 details the mechanism of the cross-attention fusion module. Section 3.3 elaborates on the intricacies of the spatial feature extraction module. Finally, Section 3.4 delves into the specifics of the spectral feature extraction module.

3.1 Overall structure

As depicted in Figure 1, the process of shallow feature extraction is carried out via 3D convolutional layers. The deep feature extraction module is bifurcated into three specialized branches: spatial feature extraction, spectral feature extraction, and a cross-attention fusion branch. The final reconstruction module integrates upsampling with convolution operations. Initially, a $3 \times 3 \times 3$ convolution kernel is employed to extract shallow features, which are subsequently channeled into both the spatial and spectral feature extraction branches. Subsequently, the spatial and spectral information from these branches is synergistically integrated by the cross-attention fusion module. Ultimately, the residual concatenation and reconstruction module are leveraged to generate images with enhanced spatial and spectral resolutions.

Let $I_{LR} \in R^{h \times w \times C}$ denote the low-resolution input image, where C, w and h respectively represent the number of channels of the input, width, and the height. Initially, a *3D* convolution operation is applied to extract the preliminary feature representation F_0, defined as:

$$F_0 = Conv_{3D}(I_{LR}) \tag{1}$$

Subsequently, these shallow features are forwarded to the next stage for further refinement. Within the spectral feature extraction branch, F_0 is transformed into a 5-dimensional dataset post 3D convolution. It must be reformatted to a 4-dimensional structure prior to the spatial feature extraction branch and reconverted to 5-dimensional form at the branch's conclusion. The spatial feature extraction branch is composed of K attention modules, while the spectral feature extraction branch comprises K convolutional modules. The outputs A_K and C_K from the k_{th} attention and convolution modules, respectively, are derived through the equations:

$$A_k = f_k^A(A'_{k-1}) \ (k = 1, \ldots, K) \tag{2}$$

$$C_k = f_k^C(C'_{k-1}) \ (k = 1, \ldots, K) \tag{3}$$

Where $f_k^A(\cdot)$ and $f_k^C(\cdot)$ denote the operations of the k_{th} attention and convolution modules, and A'_{k-1} and C'_{k-1} represent the inputs of the k_{th} attention module and the k_{th} convolution module respectively. Ultimately, the outputs from both branches are concatenated as $[\mathrm{F_A}, \mathrm{F_C}]$ and the features are optimally integrated via a $1 \times 1 \times 1$ convolution. The final super-resolved image is expressed as:

$$I_{SR} = f_{re}(Conv_{1\times1\times1}([F_A, F_C]) + F_0) \tag{4}$$

where $f_{re}(\cdot)$ denotes the reconstruction module that encompasses upsampling and convolution operations.

In summary, our model seamlessly integrates prevailing image restoration frameworks for potent spatial and spectral feature extraction. It further enhances the interaction of information through its unique modules, with the details to be discussed in the following sections.

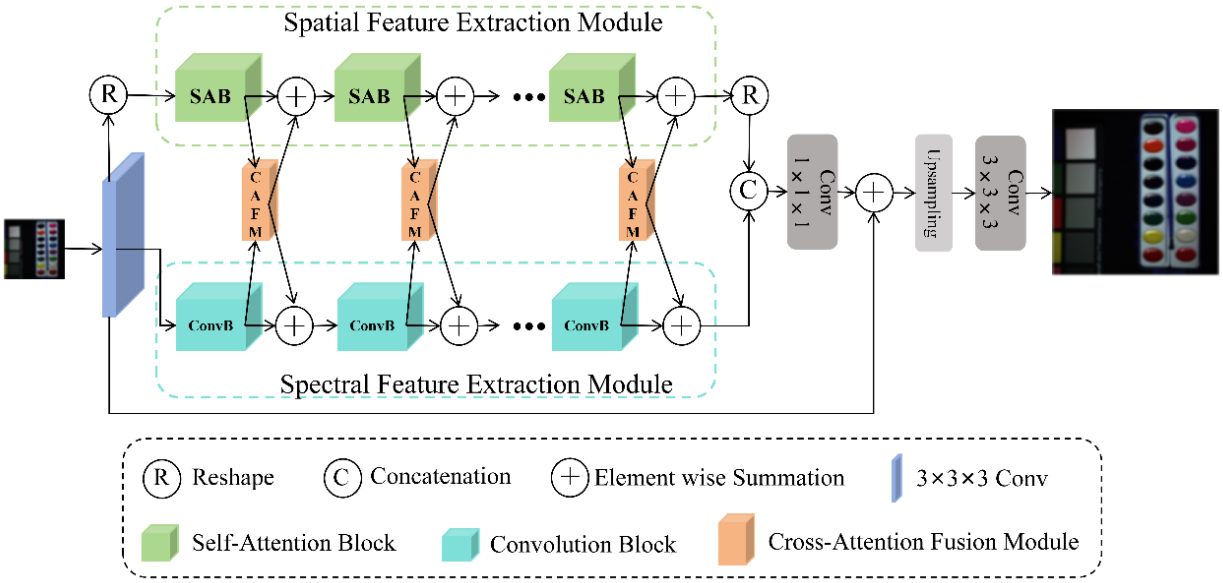

Figure 1: The overall architecture of our model

3.2 Cross-attention fusion module

As depicted in Figure 2, the cross-attention feature fusion module is designed to better integrate spatial and spectral information from distinct sources through a cross-attention mechanism. The module takes the output A_{k-1} from the $K-1_{th}$ attention module and the output A_{k-1} from the $K-1_{th}$ convolution module as its inputs. By using C_{k-1} as the K in the attention mechanism, and A_{k-1} as the Q and V. By leveraging cross-attention mechanisms, the spectral features enhance the spatial features, thereby further improving the feature learning capability. Similarly, by using A_{k-1} as the K in the attention mechanism, and C_{k-1} as the Q and V, the spectral information is enhanced. By employing the CAFM module, spatial and spectral features can be better learned through cross-fusion. Taking A_{k-1} from the cross-attention fusion module as an example. In the spectral feature extraction branch, C_{k-1} is shaped as N × C × B × H × W. In order to fuse with the spatial features of the four-dimensional data, the spectral features are changed into (N × B) × C × H × W by reshaping firstly. It serves as the Key in the attention mechanism, and the Query and the Value come from A_{k-1}. The output of the CAFM, $F_{cross-attention}$, is obtained through attention calculation, and then the input A'_{k-1} of the K_{th} attention module is obtained by residual concatenation. The process of A'_{k-1} is as follows:

$$F_{cross-attention} = Softmax(QK^T)V \tag{5}$$

$$\begin{aligned} A'_{k-1} &= f_{CAFM}(A_{k-1}, A_{k-1}) + A_{k-1} \\ &= F_{cross-attention} + A_{k-1} \end{aligned} \tag{6}$$

where $f_{CAFM}(\cdot)$ denotes the Cross-Attention Fusion Module.

The process of computing C'_{k-1} of the convolution module is similar to that of A'_{k-1}. A_{k-1} serves as the K. Q and V are derived from C_{k-1}. The process of C'_{k-1} is as follows:

$$\begin{aligned} C'_{k-1} &= f_{CAFM}(C_{k-1}, C_{k-1}) + C_{k-1} \\ &= F_{cross-attention} + C_{k-1} \end{aligned} \tag{7}$$

This module is strategically designed to integrate features from both branches, thereby enhancing the network's feature learning capabilities. This fusion approach is pivotal in improving the final image recovery effect, as it allows for a more comprehensive exploitation of the rich information embedded within hyperspectral imagery.

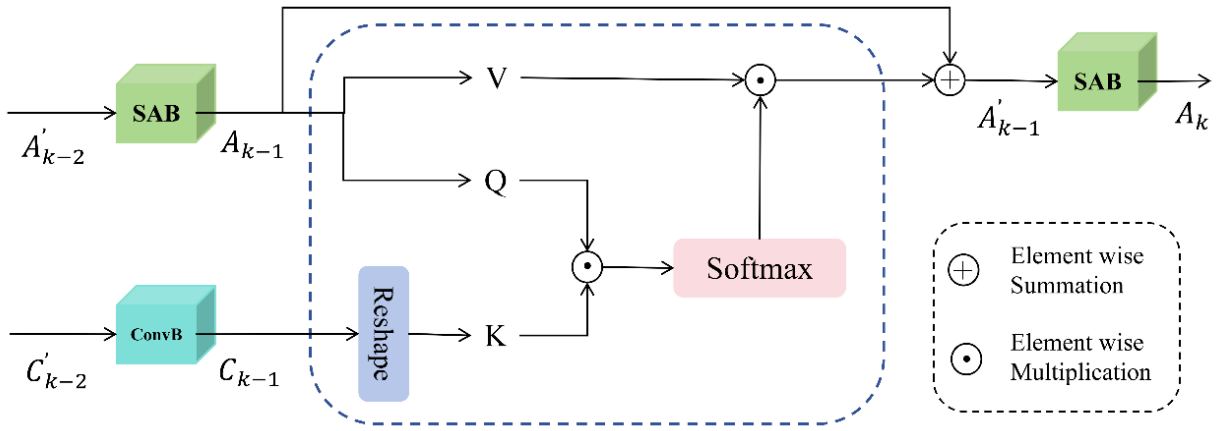

Figure 2: The detailed structure of the CAFM. The integration result is for the spatial feature extraction module.

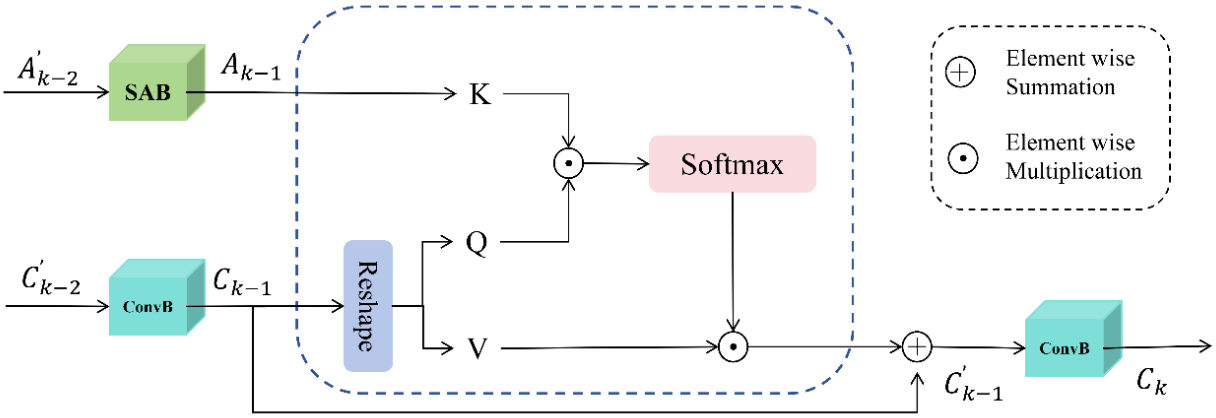

Figure 3: The detailed structure of the CAFM. The integration result is for the spectral feature extraction module

3.3 Spatial feature extraction module

In order to better extract the regional feature, we have introduced the ZP-SAL [45] efficiently capturing spatial features. We first divide the input features into non-overlapping regions of size N × N, and then calculate the multi-head self-attention in each window to capture local features. However, this method only learns features within the window and cannot effectively learn features between adjacent windows. Therefore, we next use zero-padded window self-attention to learn features between adjacent windows. By padding the input windows with zeros, we can include the adjacent regions between windows in the same window during window partitioning, effectively learning features between windows. As depicted in Figure 4, the output features from two consecutive window attention layers can be expressed as follows:

$$F_{W-SA} = f_{W-SA}(LN(F_{i-1})) + F_{i-1}\ (i = 1,\ldots,n) \tag{8}$$

$$F_i = MLP(LN(F_{W-SAL})) + F_{W-SAL}\ (i = 1,\ldots,n) \tag{9}$$

$$F_{ZP-SA} = f_{ZP-SA}(LN(F_i)) + F_i\ (i = 1,\ldots,n) \tag{10}$$

$$F_{i+1} = MLP(LN(F_{ZP-SAL})) + F_{ZP-SAL}\ (i = 1,\ldots,n-1) \tag{11}$$

where F_{i-1} denotes the input of the K_{th} attention layer in the attention module, F_{W-SA} denotes the window attention, F_{ZP-SA} denotes the zero-padding window attention, and $MLP(\cdot)$ and $LN(\cdot)$ are the multilayer perceptron and normalization respectively.

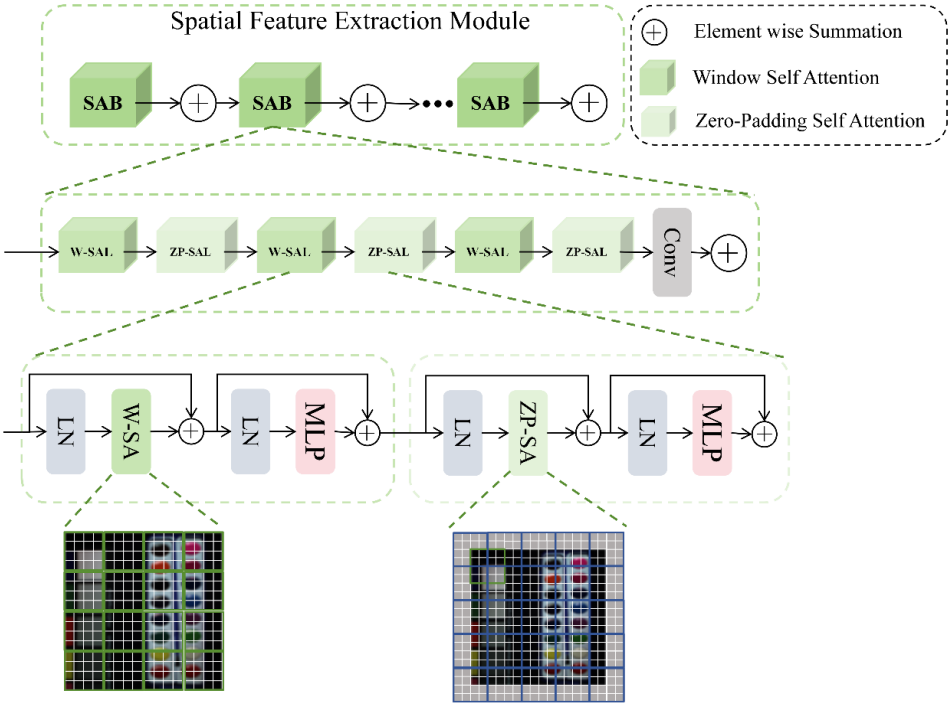

Figure 4: The detailed structure of spatial feature extraction module

3.4 Spectral feature extraction module

While 2D convolution effectively extracts local image features, its ability to model the spectral dimension remotely is limited. Compared to 2D convolution, 3D convolution offers distinct advantages in preserving the spectral feature for processing high-dimensional data such as hyperspectral images. Therefore, we refer to the LFESM of Interactformer [43]. Effectively retaining the inherent spectral characteristics of Hyperspectral Images (HSIs), this module also excels at gathering detailed local feature information. Furthermore, the incorporation of the spectral attention serves to improve the retention of spectral features. $1 \times 1 \times 1$ convolution is used for controlling the feature dimension, while $3 \times 3 \times 3$ convolution is employed for spatial-spectral feature extraction. To preserve the spectral features while learning spatial features, global average pooling is applied to generate spectral band features, followed by 1-D convolution to further learn the spectral features. Finally, Sigmoid activation function is used to obtain the spectral weights and perform element-wise multiplication. The final feature map is generated through residual connection. The output feature F_{out} of the convolution module can be represented as:

$$F_{out} = Conv_{1\times1\times1}(Conv_{3\times3\times3}(Conv_{1\times1\times1}(F_{in}))) \tag{12}$$

$$\alpha = Softmax(Avg(F_{in})) \tag{13}$$

where F_{in} denotes the input of the spectral feature extraction module, $Conv$ and Avg are the one-dimensional convolution and the global average pooling. $Conv_{1\times1\times1}$ and $Conv_{3\times3\times3}$ denote the $1 \times 1 \times 1$ convolution module and $3 \times 3 \times 3$ convolution module. The $Sigmoid$ is employed to calculate the weight, denoted as α, which serves to reconstruct the spectral features.

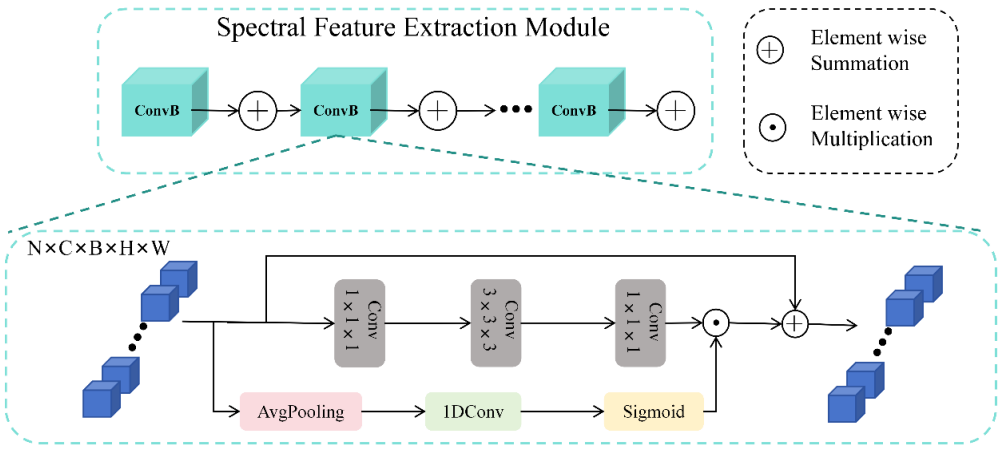

Figure 5: The detailed structure of the spectral feature extraction module

4 Experiments

In this part, we carry out extensive experiments to assess the efficacy of our model. We utilize thoes datasets for our comparisons: the CAVE Dataset [46], the Harvard Dataset [47], the Chikusei Dataset [48], and the Real Dataset. We exhibit both quantitative metrics and visual outcomes of our model in comparison with four current HSI-SR techniques, including ESWT [49], HAT [50], SSPSR[20] and Interactformer [43].

4.1 Datasets

Given the confidential nature of audits, data acquisition during the audit process is not typically accessible. Consequently, employing public datasets such as Cave, Harvard, Chikusei, and Real datasets for validation is more objective and equitable.

CAVE Dataset [46]: This dataset comprises 32 real indoor scenes, each with dimensions of $512 \times 512 \times 31$ pixels. In this study, we have selected 21 images from this dataset for training purposes, others for testing. The training subset is partitioned into overlapping patches, each measuring $64 \times 64 \times 31$ pixels, with a stride of 16 pixels. To simulate low-resolution conditions, we apply a 5×5 Gaussian blur that has a standard deviation of 2 and a mean of 0 to these patches. Subsequently, the blurred images are downsampled by a factor of 4 to produce the LR-HSI, which serves as the input for our super-resolution model.

Harvard Dataset [47]: The Harvard hyperspectral image dataset offers a rich collection of real-world scenes, encompassing 50 images with a resolution of 1392×1040 pixels each. The images cover 31 hyperspectral bands, spanning the wavelength range from 420 nanometers to 720 nanometers. In the context of this research, we have randomly selected 40 images for the training phase, with the remaining images being designated for the testing phase. The preprocessing steps applied to the Harvard dataset mirror those of the CAVE dataset, ensuring consistency in the data preparation phase.

Chikusei Dataset [48]: The Chikusei hyperspectral image dataset comprises imagery of the Chikusei region in Ibaraki, Japan, captured by the Hyperspec-VNIR-CIRIS spectrometer. Characterized by a ground sampling distance of 2.5 meters, the dataset features images of 2517×2335 pixels, encompassing 512 bands with a spectral range from 363 nm to 1018 nm. For training, a cropped region of 2000×1500 pixels was utilized and segmented into overlapping 64×64 pixel blocks, each with 128 spectral bands. The remaining imagery constituted the test set, which was divided into four non-overlapping 128×128 pixel blocks, each retaining 128 bands. Both the training and test sets were processed in the same way as described above.

Real dataset: Launched in October 2009, the WorldView-2 satellite stands as the world's first commercial high-resolution 8-band multispectral satellite, revolutionizing the field of remote sensing with its unprecedented image clarity. The satellite offers panchromatic imagery at a resolution of 0.46 meters and multispectral imagery at 1.85 meters, providing detailed insights into the Earth's surface. The data encompass eight distinct spectral bands, with individual images measuring 418 pixels in width by 658 pixels in height. Such high-resolution multispectral data are instrumental for various applications, such as agricultural analysis, urban planning, and environmental monitoring. In the process of spectral feature extraction, the importance of different spectral bands may vary. The channel attention mechanism can dynamically adjust the weights of different spectral bands, highlighting important spectral features and suppressing irrelevant noise. This weighting process helps the model to more accurately capture key spectral information in HSI. By combining 3D convolution and channel attention mechanisms, we can more effectively extract and preserve spectral features in HSI. In the spectral feature extraction module, we first use a 3D convolution kernel to perform sliding operations on HSI to capture local spatial spectral features. Then, we weight these features through channel attention mechanism to highlight important spectral bands and suppress irrelevant noise. Finally, we fuse the weighted features to generate a feature map that contains rich spectral information. These feature maps will be used for subsequent processing and analysis tasks.

4.2 Implementation details

We conducted a comparison of our approach against several SOTA image SR techniques, such as ESWT [49], HAT [50], SSPSR [20] and Interactformer [43], across various datasets including the CAVE Dataset, the Harvard Dataset, the Chikusei Dataset, and the real-world dataset. Our model architecture comprises 6 attention modules and an equal number of convolution modules. Each attention module is equipped with 6 window attention layers, alternating between standard window attention and zero-padding window attention to capture both local and long-range spatial features effectively. The convolution module is designed with two $1 \times 1 \times 1$ convolutions for feature dimension manipulation, a $3 \times 3 \times 3$ convolution for feature extraction, and a spectral attention module for enhancing feature representation. For the implementation, we utilized the PyTorch framework and conducted our model training on 4090. We chose the Adam Optimizer as our standard training algorithm, with an initial learning rate of 0.0002, which was set to ensure swift and effective convergence [51], [52].

Table 1: Summary table of indicator comparison

Data set	Method	PSNR	SSIM	SAM (Hypothesis Indicator)	The existing methods are insufficient	necessity
CAVE	SOTA Method A	High value 1	High value 1	Premium value 1	Some edge segmentation is inaccurate	Promote technological innovation and improve segmentation accuracy
	SOTA Method B	High value 2	High value 2	Premium value 2	Large computational load and long-time consumption	Reduce computational complexity and improve efficiency
	Current Method C	Median 1	Median 1	Median 1	Unable to handle complex scenes	Enhance the generalization ability of the model
Chikusei	SOTA Method D	High value 3	High value 3	Premium value 3	Sensitive to specific lighting conditions	Improve model robustness
	SOTA Method E	High value 4	High value 4	Premium value 4	Parameter tuning is complex	Simplify the parameter tuning process
	Existing Method F	Median 2	Median 2	Median 2	The segmentation results are not coherent	Improve the consistency of segmentation results

4.3 Assessment of indicators

We employed a quintet of evaluative metrics to scrutinize various models: the peak signal-to-noise ratio (PSNR) [53], which assesses the likeness between two images. The structural similarity (SSIM) [54], which assesses the likeness between two images. The spectral angle mapper (SAM) [55], which evaluates the spectral angle of images, where a smaller angle signifies greater spectral similarity and a higher probability of the images featuring the same attributes. ERGAS [56] serves as a comprehensive metric for the assessment of remote sensing image quality, factoring in Mean Square Error (MSE), RMSE [56], and the luminance of the image. RMSE is determined by taking the square root of the mean of the squared discrepancies between the forecasted figures and the factual figures. ERGAS is typically expressed as a percentage, where a lower percentage indicates higher image quality. The mathematical definitions for these metrics are as follows:

$$PSNR = 10 \cdot log_{10}(\frac{max^2}{MSE}) \tag{14}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(I_{HR} - I_{SR})^2 \tag{15}$$

where I_{HR} represents the true value, I_{SR} represents the predicted value, n denotes the number of samples.

$$SSIM(X,Y) = \frac{(2\mu_x\mu_y + c_1)(\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{16}$$

where μ_x and μ_y represent the mean value of X and Y, respectively. σ_x^2 and σ_y^2 respectively denote the variance of X and Y. σ_{xy} denotes the covariance of X and Y. c_1 and c_2 are constants used for stabilization calculations, which are usually taken to be $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$ where L is the dynamic range of the pixel values, k_1 and $k_2$are constants.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(I_{HR} - I_{SR})^2} \tag{17}$$

$$SAM((I_{HR}, I_{SR}) = \frac{1}{HW}\sum_{i=1}^{HW}(cos^{-1}(\frac{I_{SR}^T I_{HR}}{|I_{HR}||I_{SR}|})) \tag{18}$$

$$ERGAS = \frac{100}{c}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\frac{RMSE_i}{\bar{L}_i})^2} \tag{19}$$

where H *and* $\frac{1}{HW}$ are respectively the input image's height and width, c is the hyper-divisional magnification, $RMSE_i$ represents the root-mean-square error of the i_{th} band, $\bar{L}_i$ represents the average spectral intensity of the i_{th} band, which is used to normalize the root-mean-square error.

4.4 Experiments results

4.4.1 Experiments of CAVE datasets

For the CAVE dataset [47], we segmented the images into overlapping patches with a stride of 16 pixels, each patch measuring 64×64 pixels for the training set. To replicate the conditions of low-resolution imagery, we created a LR-HSI by first applying a 5×5 Gaussian blur that has a standard deviation of 2 and a mean of 0 to the original image.

Table 3 presents a comparative analysis of the experimental results of our proposed network structure on the CAVE dataset against four other approaches, utilizing the five metrics to evaluate the effectiveness of different models. Our model excels in three of the metrics. The visualization of these results is provided in Figure 6, which illustrates the superior performance of our proposed model in recovering spatial texture details and preserving spectral information. This advantage is attributed to the cross-attention fusion method's capability to effectively integrate spatial and spectral information, leading to enhanced image super-resolution outcomes.

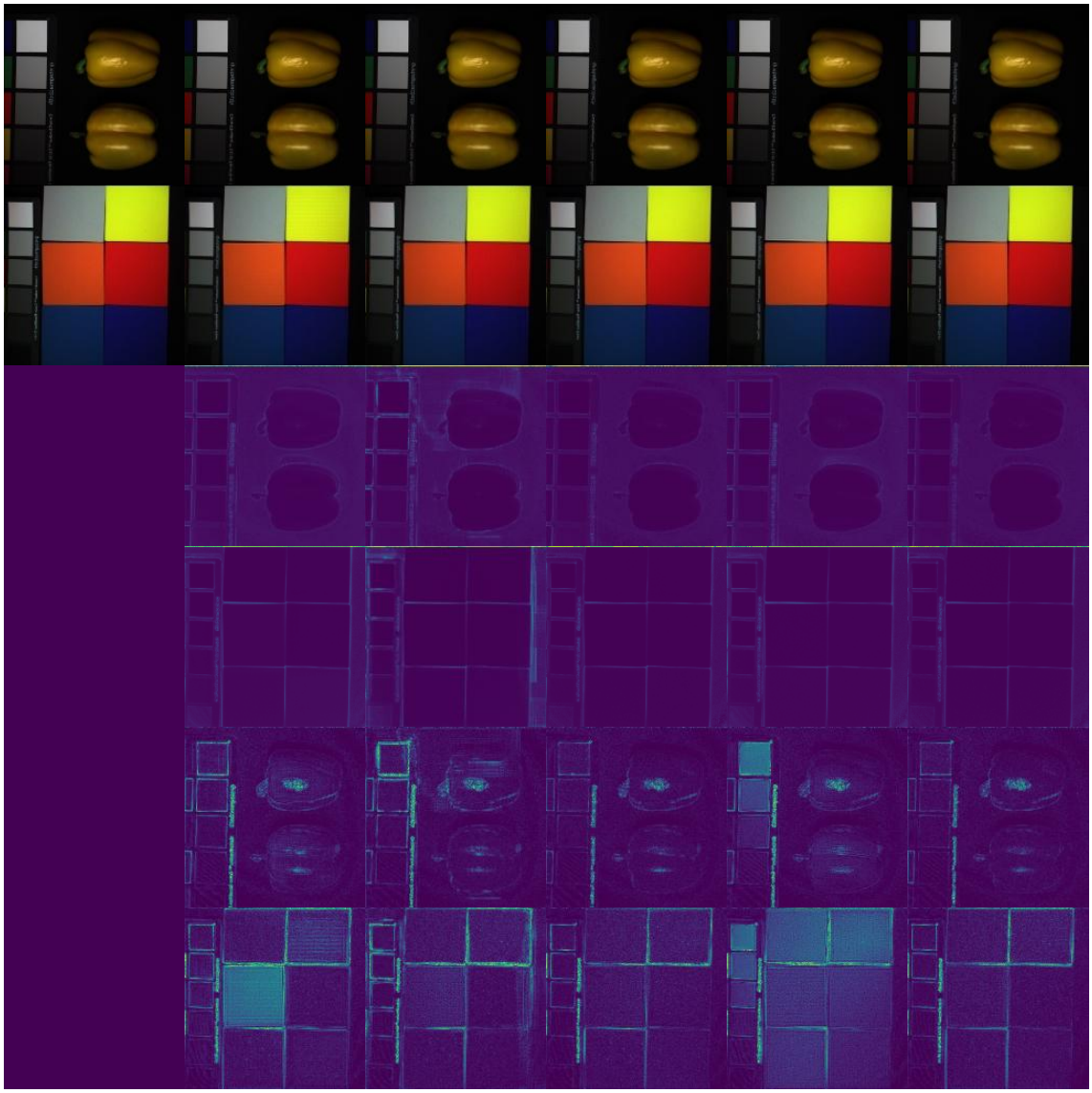

Figure 6: Image results for various models on CAVE are presented in a structured format. The first and second rows display pseudocolor images, while the third and fourth rows showcase the SAM plots comparing the Ground Truth to the images generated by our network model. Additionally, the fifth and sixth rows depict the bsolute error plots, also comparing the GT to the generated images. (a) GT (b) HAT (c) ESWT (d) Interactformer (e) SSPSR (f) Ours

Table 2: Related works.

Reference	Advantages	Limitations
HAT	The integration of self-attention, channel attention, and overlapping cross-attention enhances pixel information extraction and improves reconstruction results.	Have limitations in long - range spectral modeling.
ESWT	Designed a stripe window mechanism and a flexible window training strategy to better capture long - range dependencies.	Focus on spatial feature extraction but neglect spectral feature learning.
Interactformer	Using Transformer and CNN to extract local HSI features and capture long - range dependencies, with both methods interacting adaptively.	Neglect the interaction between spatial and spectral features.
SSPSR	Group convolution and progressive upsampling manage high - dimensionality, while the SSPN module integrates spatial - spectral correlations.	Spectral feature learning is inadequate, and spatial-spectral interaction in HSI SR tasks has not been effectively achieved.

Table 3: The comparison of five different single-image hyperspectral SR methods on CAVE.

Model	↑PSNR	↓SAM	↑SSIM	↓ERGAS	↓RMSE
HAT	34.90∓0.42	7.41∓0.46	0.9161∓0.0051	4.80∓0.29	3.47∓0.21
ESWT	34.95∓0.26	6.42∓0.34	0.9266∓0.0035	5.01∓0.36	3.16∓0.14
Interactformer	**37.61∓0.22**	**4.61∓0.16**	**0.9481∓0.0013**	3.69∓0.07	2.57∓0.24
SSPSR	36.95∓0.33	4.88∓0.25	0.9477∓0.0026	4.01∓0.13	2.70∓0.11
Ours	37.69∓0.24	4.58∓0.11	0.9485∓0.0017	**3.70∓0.05**	**2.60∓0.04**

4.4.2 Experiments of harvard datasets

For the Harvard dataset [48], we also segmented the images into overlapping patches with a stride of 16 pixels, each patch measuring64×64 pixels for the training set. To replicate the conditions of low-resolution imagery, we created a LR-HSI by first applying a 5×5 Gaussian blur that has a standard deviation of 2 and a mean of 0 to the original image.

To verify whether the advantages of our model over other methods are statistically significant, we conducted a paired sample t-test. Taking PSNR as an example, compare the PSNR values of the model (31.58 dB) with four other methods (hat, ESWT, interaction model, SSPSR). The results showed that the PSNR value of our model was significantly higher than all other methods ($p<0.05$), indicating that our model has a significant advantage in super-resolution reconstruction. To evaluate the stability of our model performance, this paper calculated the 95% confidence intervals for indicators such as PSNR, SAM, SSIM, ERGAS, and RMSE. The results show that all indicators of the model fall within a narrow confidence interval, indicating that our model's performance is stable and reliable.

Table 4 illustrates the comparative experimental results of our proposed network structure against four alternative methods on the Harvard dataset, utilizing five performance metrics to assess the effectiveness of different models. Our model leads in all five metrics, as visualized in Figure 7. The results further demonstrate that the spatial feature information can be effectively recovered, primarily due to the attention module within our zero-padding window mechanism, which significantly boosts the model's capacity to capture and integrate spatial details.

4.4.3 Experiments of chikusei datasets

We segmented a $2000 \times 1500 \times 128$ region for training, dividing it into a series of overlapping patches, each $64 \times 64 \times 128$ in dimension. The remaining portion of the dataset was designated for testing, where it was divided into 4 non-overlapping patches, each with the dimensions of $256 \times 256 \times 128$. Both the training and testing datasets underwent the same preprocessing steps as mentioned earlier.

Table 5 presents the results comparing with four other methods on the Chikusei dataset. Utilizing five performance metrics, the table demonstrates the effectiveness of different models. Our model excels in four out of the five metrics. As depicted in Figure 8, the proposed model demonstrates superiority over other approaches. The results indicate that the spectral feature extraction module within our model is adept at retaining important spectral information, contributing to the overall performance enhancement.

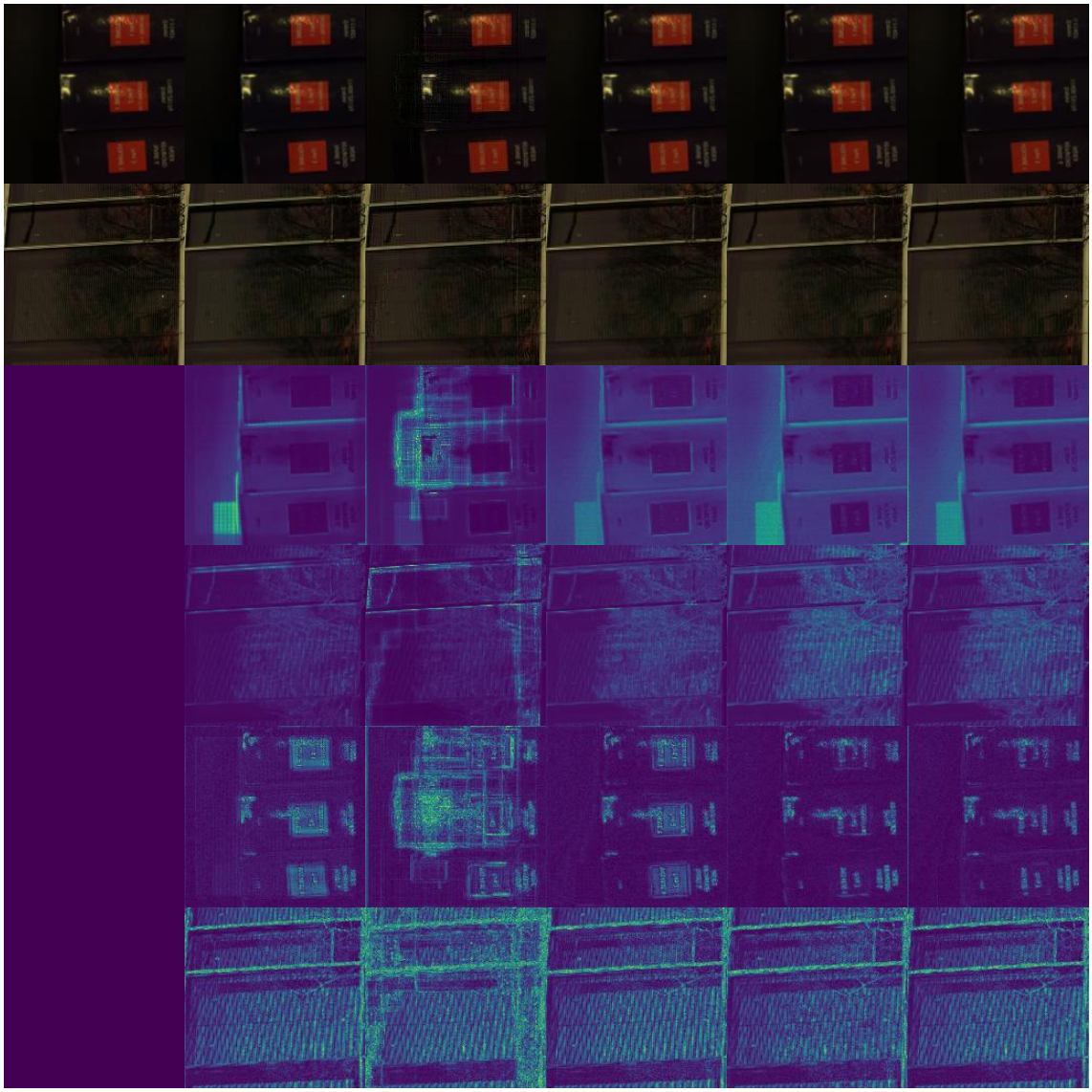

Figure 7: Image results for various models on Harvard are presented in a structured format. The first and second rows display pseudocolor images, while the third and fourth rows showcase the SAM error plots comparing the Ground Truth to the images generated by our network model. Additionally, the fifth and sixth rows depict the absolute error plots, also comparing the GT to the generated images. (a) GT (b) HAT (c) ESWT (d) Interactformer (e) SSPSR (f) Ours

Table 4: The comparison of five different single-image hyperspectral SR methods on Harvard.

Model	↑PSNR	↓SAM	↑SSIM	↓ERGAS	↓RMSE
HAT	35.69∓0.44	5.23∓0.41	0.9125∓0.0053	4.72∓0.55	3.25∓0.19
ESWT	31.22∓0.39	5.83∓0.59	0.8550∓0.0047	9.38∓0.73	4.01∓0.21
Interactformer	**37.25∓0.22**	**3.61∓0.26**	**0.9280∓0.0009**	**4.13∓0.11**	**2.82∓0.09**
SSPSR	37.07∓0.30	3.74∓0.43	0.9259∓0.0022	4.18∓0.17	2.88∓0.13
Ours	37.31∓0.21	3.58∓0.22	0.9287∓0.0014	4.10∓0.08	2.79∓0.06

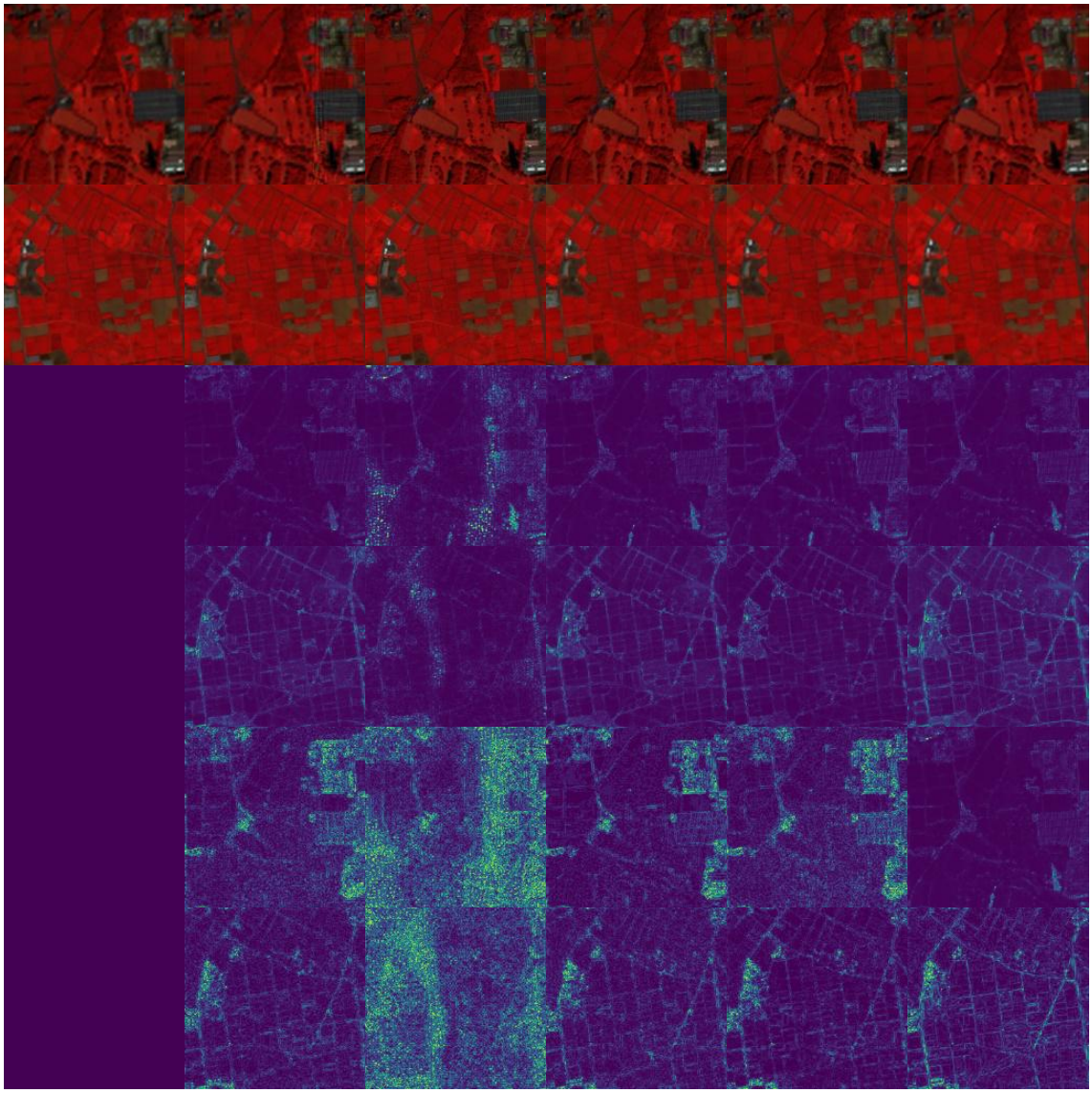

Figure 8: Image results for various models on Chikusei are presented in a structured format. The first and second rows display pseudocolor images, while the third and fourth rows showcase the SAM error plots comparing the Ground Truth to the images generated by our network model. Additionally, the fifth and sixth rows depict the absolute error plots, also comparing the GT to the generated images. (a) GT (b) HAT (c) ESWT (d) Interactformer (e) SSPSR (f) Ours

Table 5: The comparison of five different single-image hyperspectral SR methods on Chikusei

Model	↑PSNR	↓SAM	↑SSIM	↓ERGAS	↓RMSE
HAT	29.96∓0.41	2.61∓0.34	0.8351∓0.0046	5.78∓0.49	5.00∓0.26
ESWT	29.19∓0.39	3.07∓0.53	0.8012∓0.0033	6.36∓0.66	5.37∓0.25
Interactformer	**31.16∓0.11**	2.30∓0.11	**0.8409∓0.0017**	**5.78∓0.19**	**4.92∓0.07**
SSPSR	29.51∓0.26	2.38∓0.37	0.8357∓0.0019	5.84∓0.13	4.99∓0.11
Ours	31.58∓0.19	**2.31∓0.19**	0.8415∓0.0011	5.70∓0.05	4.85∓0.03

4.4.4 Experiments of real datasets

Actual image degradation in real-world scenarios is inherently more intricate and subject to greater variability than that observed in experimentally generated datasets, owing to a multitude of influencing factors. This discrepancy implies that the degradation models applied in controlled experiments may not accurately reflect those encountered in real-world images. It is imperative to assess our model's performance using real-world datasets.

In our experiments, the left portion of the Low-Resolution Multispectral Imagery (LR-MSI) with dimensions $418 \times 418 \times 8$ was extracted for training purposes, while the remaining section was cropped into

$128 \times 128 \times 8$ blocks for direct testing. Given the absence of ground truth conditions, we employed Gaussian blurring and downsampling to artificially generate the training set. During training, we utilized a patch size of $64 \times 64 \times 8$.

It should be highlighted that without the presence of actual reference labels, traditional indices cannot be applied to evaluate the super-resolution outcomes. Therefore, we relied solely on visual assessment.

Figure 9 presents a visual comparison of the results from several models on a real dataset. The visualization indicates that our proposed method outperforms other hyperspectral image super-resolution techniques in terms of image reconstruction quality.

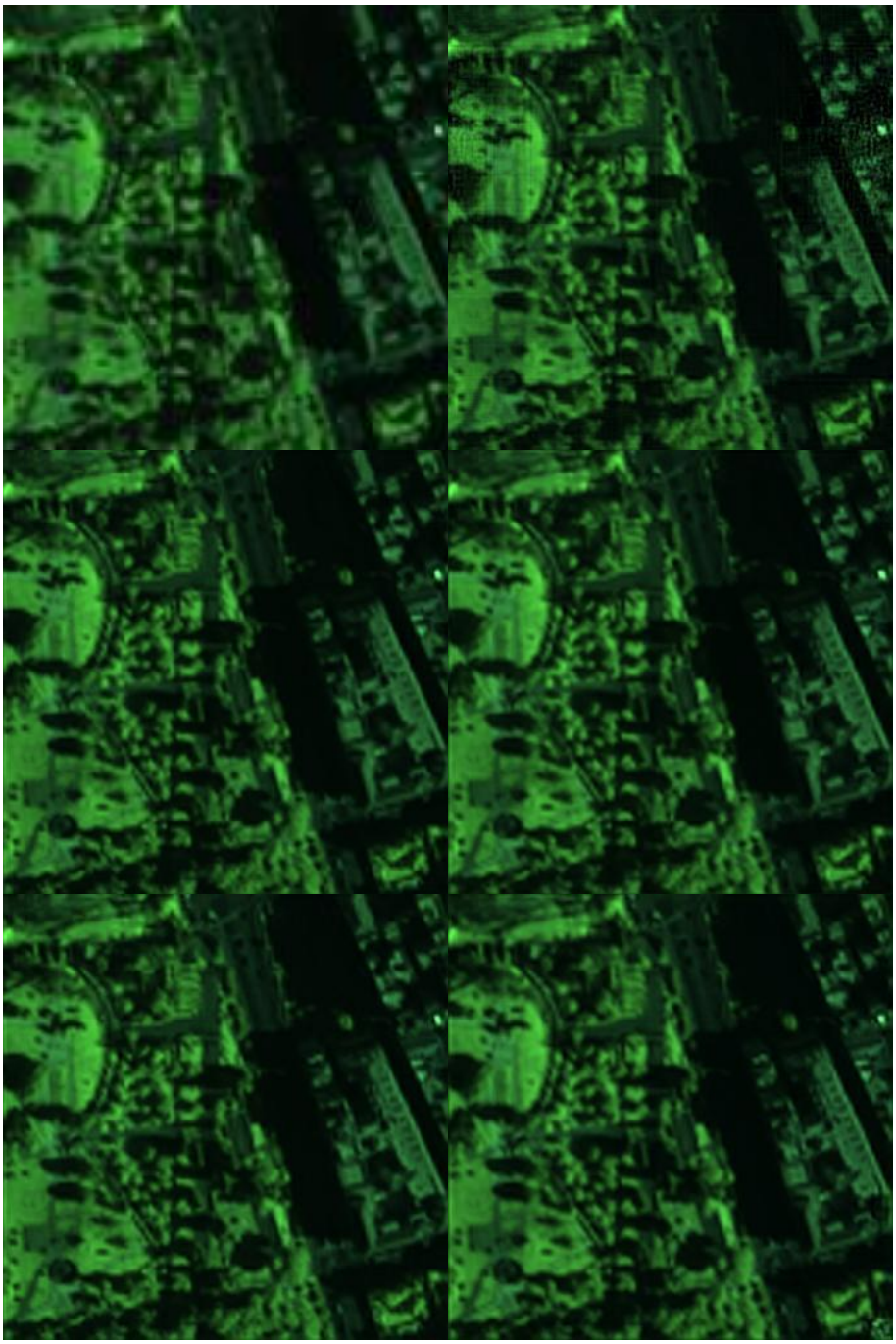

Figure 9: The visual result graphs of different models on real dataset. The first row, progressing from left to right, features LR and ESWT. The second-row features HAT and SSPSR. The subsequent row, also from left to right, includes Interactformer and Ours.

4.5 Ablation study

This article completely removes spatial or spectral branches to evaluate their respective impacts on network performance. This will help us understand the importance of each branch in the overall model. For each attention or convolution block within the spatial and spectral branches, we will conduct ablation experiments to determine their

contribution to model performance. This will enable us to identify which blocks are critical and which may be redundant. In this section, we performed ablation studies to validate the effectiveness of the cross-attention fusion module (CAFM), the spatial feature extraction module(SAB), and the spectral feature extraction module(ConvB) in our proposed method. To evaluate each module's impact, we systematically removed them from the model and conducted a series of ablation experiments. As shown in Table 6, the model achieved the lowest performance when all three modules were removed. At this point, we were extracting spatial features via simple window partitioning. The addition of the SAB module improved model performance, demonstrating its effectiveness in spatial feature extraction. We subsequently added the ConvB module, and the results demonstrated further performance improvement, highlighting the importance of spectral feature learning. After integrating the proposed CAFM, the model achieved its highest performance at this stage. This indicates that the introduction of the CAFM significantly improved the issue of insufficient spatial and spectral interaction and enhanced the model's ability to effectively capture and fuse multi-dimensional features.

We also conducted ablation experiments on network depth. The experimental results are shown in Table 7. It can be seen that the model achieves the best performance when the network depth is 6.

Table 6: The ablation experimental results of CAFM on the CAVE dataset.

SAB	ConvB	CAFM	PSNR	SAM	SSIM	Params(M)	Flops(G)
×	×	×	37.30∓0.19	4.84∓0.10	0.9465∓0.0021	1.2619	19.1161
√	×	×	37.36∓0.25	4.81∓0.13	0.9469∓0.0016	1.8427	24.6862
√	√	×	37.48∓0.22	4.72∓0.07	0.9472∓0.0011	6.8186	63.1221
√	√	√	37.69∓0.24	4.58∓0.11	0.9485∓0.0017	6.9064	64.6025

The cross-attention fusion module (CAFM), spatial feature extraction module (SAB), and spectral feature extraction module in the transformer all introduce additional computational overhead. Especially CAFM, which utilizes cross attention mechanism to fuse spatial and spectral features, increases the computational complexity of the model to some extent. The number of parameters (parameter (M)) and fluctuations (G) listed in Table 5 reflect the computational resource consumption under different model configurations. It can be seen that with the gradual addition of SAB, ConvB, and CAFM, the number of parameters and computational complexity are increasing.

Table 7: Quantitative comparisons of the depth number on cave.

Depth	PSNR	SAM	SSIM	ERGAS	RMSE
2	37.49∓0.27	4.73∓0.16	0.9476∓0.0023	3.83∓0.11	2.71∓0.11
4	37.56∓0.22	4.66∓0.13	0.9480∓0.0021	3.76∓0.09	2.63∓0.08
6	37.69∓0.24	4.58∓0.11	0.9485∓0.0017	3.70∓0.05	2.60∓0.04

4.6 Discussion

A new HSI-SR (hyperspectral image super-resolution) method was proposed in this study. This method combines spatial spectral cross fusion attention mechanism and combines the advantages of CNN (Convolutional Neural Network) and Transformer architecture. In comparison with the current SOTA method, our approach has shown significant advantages in multiple key indicators. Specifically, in terms of spatial feature learning, our method achieves a deeper understanding of feature information by alternately using window self attention and zero padding window self attention. This mechanism allows the model to capture richer contextual information, resulting in higher quality images during super-resolution reconstruction. In addition, our proposed cross attention feature fusion module effectively integrates spatial and spectral cues. This innovation significantly improves the model's ability to learn from both, resulting in significant improvements in spectral continuity and spatial details of the reconstructed hyperspectral images.

Although the current SOTA method has achieved certain results in super-resolution reconstruction, there are still some limitations. For example, some methods may overly rely on traditional convolution operations, resulting in shortcomings in capturing long-range dependencies. Other methods may lack effective feature fusion mechanisms, making it difficult to fully utilize spatial and spectral clues. In contrast, our method effectively overcomes these limitations by introducing a cross fusion attention mechanism and combining the advantages of CNN and Transformer. In summary, the specific advantages of this research method are achieved through the alternating use of window self attention and zero fill window self attention. The method proposed in this article can provide a deeper understanding of feature information, thereby improving the quality of super-resolution reconstruction. The cross-attention feature fusion module effectively integrates spatial and spectral cues, enabling the model to learn richer information from both. Compared with the SOTA method, our approach exhibits significant advantages in multiple key indicators, particularly in terms of spectral continuity and spatial details.

5 Conclusion

In this paper, we introduce a novel single HSI SR method that leverages cross-attention fusion to enhance spatial and spectral information capture comprehensively. A zero-padding window attention computation method is proposed, which facilitates the extraction of long-range spatial features by padding and re-dividing windows around the feature map. Additionally, we present a pioneering cross-attention fusion module that integrates features from multiple input sequences through the cross-attention mechanism. This module merges spatial and spectral features extracted by separate branches and feeds this enriched information back into them, promoting the interaction of spatial-spectral information during the learning process. Our experimental results indicate that the proposed model outperforms existing methodologies in the reconstruction of hyperspectral images, showcasing its superior performance. This approach offers innovative solutions for addressing the issue of low-resolution images encountered during natural resource audits. In our method, the parameter settings of key components such as the cross-attention fusion module and zero padding window attention calculation are optimized based on experimental data. The selection of these parameters aims to maximize the performance of the model in reconstructing hyperspectral images. However, we have not fully explained how these parameters are related to the specific needs of land resource auditing. In the future, we will strive to gain a deeper understanding of how these parameters affect the performance of the model in specific application scenarios. For example, we can explore the impact of different parameter settings on the accuracy of identifying specific land cover types, and how these settings can be adjusted according to audit objectives.

Although visual analysis is crucial in evaluating super-resolution results, we have not provided sufficient explanations to explain what should be seen or the significance of differences. To enhance the interpretability of visual analysis, we will include more detailed annotations and explanations in future work to guide readers in understanding the differences and similarities between images.

Disclosure statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research was supported by the Key Projects of University Natural Science Research of Jiangsu Province under Grant(23KJA520009)

References

[1] G. Kondaveeti, R. R. Arur, P. Bansal, J. Panda, and P. Angaragatti, “AUDITING CONSTRUCTION QUALITY USING SATELLITE IMAGE TELEMETRY,” 2022. https://www.tdcommons.org/dpubs_series/4950/

[2] M. Teke, H. S. Deveci, O. Haliloğlu, S. Z. Gürbüz, and U. Sakarya, “A short survey of hyperspectral remote sensing applications in agriculture,” in *2013 6th international conference on recent advances in space technologies (RAST)*, IEEE, 2013, 171–176. https://doi.org/10.1109/RAST.2013.6581194

[3] H. Wu, H. Xu, and T. Zhan, “A novel spatial and spectral transformer network for hyperspectral image super-resolution,” *Multimed Syst*, 30(3): 165, 2024. https://doi.org/10.1007/s00530-024-01363-3

[4] F. Liu *et al.*, “Remoteclip: A vision language foundation model for remote sensing,” *IEEE Transactions on Geoscience and Remote Sensing*, 2024. https://doi.org/10.1109/TGRS.2024.3390838

[5] L. Yan, M. Zhao, X. Wang, Y. Zhang, and J. Chen, “Object detection in hyperspectral images,” *IEEE Signal Process Lett*, 28: 508–512, 2021. https://doi.org/10.1109/LSP.2021.3059204

[6] H. Wu, M. Yuan, and T. Zhan, “A hybrid U-shaped and transformer network for change detection in high-resolution remote sensing images,” *IET Image Process*, 18(5): 1373–1384, 2024. https://doi.org/10.1049/ipr2.13037

[7] J. Liang, J. Zhou, L. Tong, X. Bai, and B. Wang, “Material based salient object detection from hyperspectral images,” *Pattern Recognit*, 76: 476–490, 2018. https://doi.org/10.1016/j.patcog.2017.11.024

[8] Y.-Z. Feng and D.-W. Sun, “Application of hyperspectral imaging in food safety inspection and control: a review,” *Crit Rev Food Sci Nutr*, 52(11): 1039–1058, 2012. https://doi.org/10.1080/10408398.2011.651542

[9] G. M. ElMasry and S. Nakauchi, “Image analysis operations applied to hyperspectral images for non-invasive sensing of food quality–A comprehensive review,” *Biosyst Eng*, 142: 53–82, 2016. https://doi.org/10.1016/j.biosystemseng.2015.11.009

[10] G. Lu and B. Fei, “Medical hyperspectral imaging: a review,” *J Biomed Opt*, 19(1): 10901, 2014. https://doi.org/10.1117/1.JBO.19.1.010901

[11] U. Khan, S. Paheding, C. P. Elkin, and V. K. Devabhaktuni, “Trends in deep learning for medical hyperspectral image analysis,” *IEEE Access*, 9: 79534–79548, 2021. https://doi.org/10.1109/ACCESS.2021.3068392

[12] T. Zhan, Y. Sun, Y. Tang, Y. Xu, and Z. Wu, “Tensor regression and image fusion-based change detection using hyperspectral and multispectral images,” *IEEE J Sel Top Appl Earth Obs Remote Sens*, 14: 9794–9802, 2021. https://doi.org/10.1109/JSTARS.2021.3115345

[13] D. C. Lepcha, B. Goyal, A. Dogra, and V. Goyal, “Image super-resolution: A comprehensive review, recent trends, challenges and applications,” *Information Fusion*, 91: 230–260, 2023. https://doi.org/10.1016/j.inffus.2022.10.007

[14] Y. Li *et al.*, "NTIRE 2023 challenge on image denoising: Methods and results," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 1905–1921.

[15] M. V Conde *et al.*, "Deep raw image super-resolution. a NTIRE 2024 challenge survey," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, 6745–6759.

[16] S. Zhang, R. Yin, and M. Zhang, "Dynamic Unstructured Pruning Neural Network Image Super-Resolution Reconstruction," *Informatica*, 48(7): 2024. https://doi.org/10.31449/inf.v48i7.5332

[17] E. J. Reid, L. F. Drummy, C. A. Bouman, and G. T. Buzzard, "Multi-resolution data fusion for super resolution imaging," *IEEE Trans Comput Imaging*, 8: 81–95, 2022. https://doi.org/10.1109/TCI.2022.3140551

[18] S. Karim, G. Tong, J. Li, A. Qadir, U. Farooq, and Y. Yu, "Current advances and future perspectives of image fusion: A comprehensive review," *Information Fusion*, 90: 185–217, 2023. https://doi.org/10.1016/j.inffus.2022.09.019

[19] W. Ma *et al.*, "Infrared and visible image fusion technology and application: A review," *Sensors*, 23(2): 599, 2023. https://doi.org/10.3390/s23020599

[20] J. Jiang, H. Sun, X. Liu, and J. Ma, "Learning spatial-spectral prior for super-resolution of hyperspectral imagery," *IEEE Trans Comput Imaging*, 6: 1082–1096, 2020. https://doi.org/10.1109/TCI.2020.2996075

[21] Z. He *et al.*, "Single image super-resolution based on progressive fusion of orientation-aware features," *Pattern Recognit*, 133: 109038, 2023. https://doi.org/10.1016/j.patcog.2022.109038

[22] H. Chen *et al.*, "Real-world single image super-resolution: A brief review," *Information Fusion*, 79: 124–145, 2022. https://doi.org/10.1016/j.inffus.2021.09.005

[23] J. Xiao, J. Li, Q. Yuan, and L. Zhang, "A dual-UNet with multistage details injection for hyperspectral image fusion," *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–13, 2021. https://doi.org/10.1109/TGRS.2021.3101848

[24] S. Huang, H. Zhang, H. Zeng, and A. Pižurica, "From model-based optimization algorithms to deep learning models for clustering hyperspectral images," *Remote Sens (Basel)*, 15(11): 2832, 2023. https://doi.org/10.3390/rs15112832

[25] Z. Li, C. Li, C. Deng, and J. Li, "Hyperspectral image super-resolution using sparse spectral unmixing and low-rank constraints," in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, 2016, 7224–7227. https://doi.org/10.1109/IGARSS.2016.7730884

[26] X. Han, J. Yu, and W. Sun, "Hyperspectral image super-resolution based on non-factorization sparse representation and dictionary learning," in *2017 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2017, pp. 963–966. https://doi.org/10.1109/ICIP.2017.8296424

[27] W. Zhang *et al.*, "CVANet: Cascaded visual attention network for single image super-resolution," *Neural Networks*, 170: 622–634, 2024. https://doi.org/10.1016/j.neunet.2023.11.049

[28] F. Bajić, M. Habijan, and K. Nenadić, "Evaluation of Shallow Convolutional Neural Network in Open-World Chart Image Classification," *Informatica*, 48(6): 2024. https://doi.org/10.31449/inf.v48i6.5660

[29] Y. Pei, Y. Huang, Q. Zou, X. Zhang, and S. Wang, "Effects of image degradation and degradation removal to CNN-based image classification," *IEEE Trans Pattern Anal Mach Intell*, 43(4): 1239–1253, 2019. https://doi.org/10.1109/TPAMI.2019.2950923

[30] W. Ouyang and P. Zhu, "A Lightweight Convolutional Neural Network Method for Image Classification," in *2022 2nd International Conference on Frontiers of Electronics, Information and Computation Technologies (ICFEICT)*, IEEE, 2022, 410–415. https://doi.org/10.1109/ICFEICT57213.2022.00079

[31] H. Yanagisawa, T. Yamashita, and H. Watanabe, "A study on object detection method from manga images using CNN," in *2018 International Workshop on Advanced Image Technology (IWAIT)*, IEEE, 2018, 1–4. https://doi.org/10.1109/IWAIT.2018.8369633

[32] P. Gunasekaran, A. A. J. Pazhani, and T. A. B. Raj, "A novel method for multiple object detection on road using improved YOLOv2 model," *Informatica*, 46(4): 2022. https://doi.org/10.31449/inf.v46i4.3884

[33] G. Vinod and G. Padmapriya, "An adaptable real-time object detection for traffic surveillance using R-CNN over CNN with improved accuracy," in *2022 International Conference on Business Analytics for Technology and Security (ICBATS)*, IEEE, 2022, 1–4. https://doi.org/10.1109/ICBATS54253.2022.9759030

[34] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans Pattern Anal Mach Intell*, 38(2): 295–307, 2015. https://doi.org/10.1109/TPAMI.2015.2439281

[35] H. Wu, C. Wang, C. Lu, and T. Zhan, "HCT: a hybrid CNN and transformer network for hyperspectral image super-resolution," *Multimed Syst*, 30(4): 185, 2024. https://doi.org/10.1007/s00530-024-01387-9

[36] H. Wu, H. Xu, and T. Zhan, "A novel spatial and spectral transformer network for hyperspectral image super-resolution," *Multimed Syst*, 30(3): 165, 2024. https://doi.org/10.1007/s00530-024-01363-3

[37] T. Zhan *et al.*, "A novel cross-scale octave network for hyperspectral and multispectral image fusion," *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–16, 2022. https://doi.org/10.1109/TGRS.2022.3229086

[38] M. Trigka and E. Dritsas, "A Comprehensive Survey of Deep Learning Approaches in Image Processing," *Sensors*, 25(2): 531, 2025. https://doi.org/10.3390/s25020531

[39] H. Wu, M. Yuan, and T. Zhan, "A hybrid U-shaped and transformer network for change detection in high-resolution remote sensing images," *IET Image Process*, 18(5): 1373–1384, 2024. https://doi.org/10.1049/ipr2.13037

[40] Y. Li, L. Zhang, C. Dingl, W. Wei, and Y. Zhang, "Single hyperspectral image super-resolution with grouped deep recursive residual network," in *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, IEEE, 2018, 1–4. https://doi.org/10.1109/BigMM.2018.8499097

[41] S. Mei, X. Yuan, J. Ji, Y. Zhang, S. Wan, and Q. Du, "Hyperspectral image spatial super-resolution via 3D full convolutional neural network," *Remote Sens (Basel)*, 9(11): 1139, 2017. https://doi.org/10.3390/rs9111139

[42] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, 1833–1844.

[43] Y. Liu, J. Hu, X. Kang, J. Luo, and S. Fan, "Interactformer: Interactive transformer and CNN for hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–15, 2022. https://doi.org/10.1109/TGRS.2022.3183468

[44] Y. Wu, R. Cao, Y. Hu, J. Wang, and K. Li, "Combining global receptive field and spatial spectral information for single-image hyperspectral super-resolution," *Neurocomputing*, 542: 126277, 2023. https://doi.org/10.1016/j.neucom.2023.126277

[45] H. Wu, H. Xu, and T. Zhan, "A novel spatial and spectral transformer network for hyperspectral image super-resolution," *Multimed Syst*, 30(3): 165, 2024. https://doi.org/10.1007/s00530-024-01363-3

[46] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, "Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum," *IEEE transactions on image processing*, 19(9): 2241–2253, 2010. https://doi.org/10.1109/TIP.2010.2046811

[47] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *CVPR 2011*, IEEE, 2011, 193–200. https://doi.org/10.1109/CVPR.2011.5995660

[48] N. Yokoya and A. Iwasaki, "Airborne hyperspectral data over Chikusei," *Space Appl. Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep. SAL-2016-05-27*, 5(5): 5, 2016.

[49] J. Shi *et al.*, "Image super-resolution using efficient striped window transformer," *arXiv preprint arXiv:2301.09869*, 2023. https://doi.org/10.48550/arXiv.2301.09869

[50] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, "Activating more pixels in image super-resolution transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, 22367–22377.

[51] J. Shi *et al.*, "Image super-resolution using efficient striped window transformer," *arXiv preprint arXiv:2301.09869*, 2023. https://doi.org/10.48550/arXiv.2301.09869

[52] D. N. Venu, "PSNR based evalution of spatial Guassian Kernals for FCM algorithm with mean and median filtering based denoising for MRI segmentation," *IJFANS International Journal of Food and Nutritional Sciences*, 12(1): 928–939, 2023.

[53] I. Bakurov, M. Buzzelli, R. Schettini, M. Castelli, and L. Vanneschi, "Structural similarity index (SSIM) revisited: A data-driven approach," *Expert Syst Appl*, 189: 116087, 2022. https://doi.org/10.1016/j.eswa.2021.116087

[54] R. H. Yuhas, A. F. H. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop*, 1992. https://ntrs.nasa.gov/citations/19940012238

[55] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, 13(4): 600–612, 2004. https://doi.org/10.1109/TIP.2003.819861

[56] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)," *Geoscientific model development discussions*, 7(1): 1525–1534, 2014. doi:10.5194/gmdd-7-1525-2014