IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium | 979-8-3503-2010-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/IGARSS52108.2023.10281753

A SWIN TRANSFORMER-BASED FUSION APPROACH FOR HYPERSPECTRAL IMAGE SUPER-RESOLUTION

Yuchao Yang, Yulei Wang, Enyu Zhao, Meiping Song and Qiang Zhang

Center for Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian 116026, China

ABSTRACT

Hyperspectral image (HSI) has attracted much attention because of its rich spectral information. However, due to the limitation of imaging hardware conditions, it is often difficult to directly obtain a high spatial resolution hyperspectral image (HR-HSI). To improve the resolution, it is an economical and effective method to fuse the hyperspectral image with the high spatial resolution multi-spectral image (HR-MSI) collected from the same scene. In recent years, with the development of deep learning, the convolutional neural network (CNN) based models have been applied to solve the super-resolution reconstruction of hyperspectral images. However, limited by the convolution kernel size, the receptive field of CNN is relatively small with more attention to the local information of the image. In order to solve this problem, this paper proposes a Swin Transformer based super-resolution reconstruction (STSR) network for hyperspectral images. Specifically, Swin Transformer structure is innovatively used in STSR as the skeleton of the network, where the Swin Transformer residuals are used to extract the global spatial feature information in the image. In addition, in order to retain the spectral details in the process of super-resolution reconstruction, a spectral attention module is introduced to preserve the original spectral information. The experimental results show that the high-resolution hyperspectral images fused by the proposed STSR method are superior to the comparison method in terms of vision and quality, which proves the superiority of this method.

Index Terms—Hyperspectral image, spectral attention, swin-transformer, super-resolution

1. INTRODUCTION

Hyperspectral imaging system can collect surface information with hundreds of continuous bands simultaneously, and obtain a set of spectral images of the same scene. Compared with traditional natural or multispectral images, the main advantage of hyperspectral images is that they have richer spectral information of ground objects, which is conducive to the accurate distinction and identification of things in image scenes, so they are widely used in geological exploration, target recognition, medical diagnosis, and other fields [1]. However, due to the limitations of incident energy and hardware conditions, the spatial resolution of hyperspectral images is generally low, which also significantly affects its applications [2]. A practical solution is to reconstruct hyperspectral images by fusing higher-resolution multispectral images of the same scene. This process is normally called hyperspectral image super-resolution reconstruction.

Hyperspectral image super-resolution reconstruction (HSI-SR) is an ill-posed problem. Traditional methods try to make use of correlation between spectral bands to manually construct different prior knowledge (self-similarity, sparsity, and low rank, et al [3-5]) as a regularizer to solve this problem, which have achieved excellent performance since these used priors have been closer and closer to the basic characteristics of the data. However, these manually constructed priors have limitations and need to be readjusted when facing different data sets, thus affecting the quality of reconstructed images.

In recent years, deep learning technologies, especially CNN, have been developing rapidly. Deep learning has become a promising method to deal with the super-resolution problem of hyperspectral images. Compared with the traditional method based on manually designed priors, the CNN-based deep learning method is data-driven, and the network will learn corresponding priors according to the characteristics of the data set itself. However, the learning ability of CNN is limited. It requires deep networks when learning global features of images, and convolution is ineffective for long-distance-dependent modeling. In contrast, self-attention mechanisms proposed in the field of natural language processing (NLP) are more effective for combining global features at an early layer, especially the Transformer network, which can be used as an alternative to CNN to capture global interactions between contexts. Vit-Transformer [6] is the first attempt to apply Transformer architecture in the field of computer vision (CV), but its global self-attention mechanism has quadratic computational complexity for the input image size with a

high usage of GPU memory. On this basis, Liu et al [7]. proposed a Swin Transformer structure with a hierarchical design including shift window operation, which extends the applicability of Transformer and makes it a universal backbone of CV.

Inspired by Swin Transformer and the attention mechanism, this paper constructs a hyperspectral image super-resolution reconstruction network based on Swin Transformer. Specifically, on the one hand, convolution and several Swin Transformer residual blocks are respectively used to learn the shallow and deep features in the input image space; while on the other hand, the spectral attention module is introduced to excavate the spectral features between adjacent bands of hyperspectral images to guide image reconstruction, which can better retain the original spectral information of images. Besides, long and short skip links are added to the network, which makes the transmission of network information flow more flexible and enhances the robustness of the network.

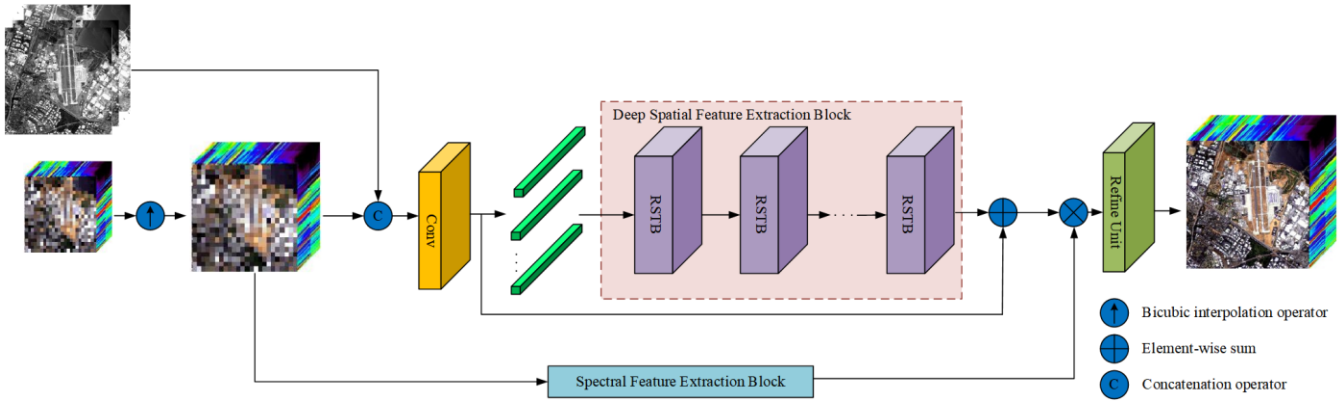

Figure 1. The overall structure of the proposed STSR algorithm.

2. METHODOLOGY

2.1. Network structure

In HSI-SR tasks, larger receptive fields can normally lead to better reconstruction results. However, it is not capable of long-range modeling due to the limitations of standard convolution operations in CNN architectures. In this paper, a Swin-Transformer architecture is used instead of CNN to capture the global spatial features of images. At the same time, in the process of image reconstruction, it is also essential to learn the spectral characteristics of the image, which helps to ensure that the spectrum of the reconstructed image is not distorted. Based on the above analysis, this paper proposes a new Swin Transformer based super-resolution reconstruction (STSR) network for hyperspectral images, as shown in Figure 1. The algorithm consists of three parts: shallow feature extraction, deep feature extraction, and image reconstruction module.

Let $Y \in \mathbb{R}^{h\times w\times L}$ and $Z \in \mathbb{R}^{H\times W\times l}$ represent the LR-HSI and HR-MSI of the same scene observed, where $W(w)$ and $H(h)$ represent the width and height of spatial dimensions, and $L(l)$ is the number of bands. The goal of super-resolution reconstruction is to estimate the HR-HSI $X \in \mathbb{R}^{H\times W\times L}$ with both high spatial and high spectral resolution from these two images.

Shallow feature extraction: Since the spatial sizes of LR-HSI and HR-MSI images are different, bicubic interpolation is firstly used to up-sample the LR-HSI data to the same size as HR-MSI, thus that the network can learn the features of both images at the same time. The MSI is then embedded into the up-sampled HSI in the spectral dimension. Finally, a simple 3×3 convolution layer is used to extract shallow features from the input fused image.

Deep feature extraction: The Residual Swin Transformer Block has been successfully used in the super-resolution reconstruction of RGB images for the first time since the SwinIR method is proposed [9], and excellent reconstruction results have been obtained. However, since HSI is a 3D data cube, both spatial and spectral self-similarity are very important in the process of super-resolution reconstruction. Therefore, this module is divided into two parts: deep spatial feature extraction block and spectral feature extraction block.

The deep spatial feature extraction block consists of N Swin Transformer residual blocks to learn the global spatial information of the image. Transformer was originally applied to NLP, and its input is all text words with a fixed scale. When it is applied to CV, there is a problem that the image resolution is much larger than the text words. Compared with traditional images, hyperspectral images have a higher spectral resolution, and each pixel of the image can be regarded as a vector with a fixed scale of band number. As a result, the global relationships between pixels can be learned by Swin Transformer residual blocks by inputting the data in the form of pixel by pixel instead of dividing the image into image blocks of uniform sizes. At the same time, the cost of calculating the global self-

attention of the image can be reduced by using the shifted window structure in the Swin Transformer residual block.

The structure of the spectral feature extraction block is shown in Figure 2. Firstly, the input global spatial information is compressed through the maximum pooling layer and the average pooling layer, and then feature learning is performed on the channel dimension through the weight-sharing multilayer perceptron (MLP). Finally, through the sigmoid layer, the spectral information of the channel is converted into a weight coefficient, which is used to measure the importance of different channels.

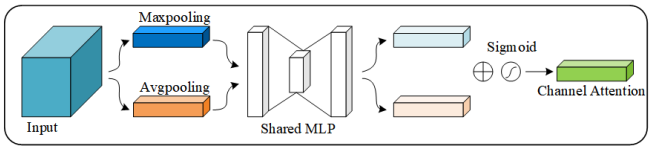

Figure 2. Schematic of the spectral feature extraction block

Image reconstruction module: In this stage, the shallow and deep spatial feature information is firstly fused, where the shallow features mainly contain low-frequency information, and the deep features mainly focus on restoring lost high-frequency information. Through global skip connections, the robustness of the network can be enhanced. Furthermore, the difficulty of training can be reduced since that the Swin Transformer residual block is more focused on mining high-frequency information. Then use spectral attention to reduce the spectral distortion in the process of learning spatial features, finally reduce the number of feature channels to the number of spectral bands through the image reconstruction block, and generate the final reconstruction results.

2.2. Loss function

In the reconstruction process, the most crucial part is restoring the high-frequency details lost in the original image. The mean absolute error (MAE) can find minor errors easier with a better convergency of the network. Therefore, this paper uses the MAE between reconstructed images and ground truth to control the learning of spatial information, as shown in equation (1).

$$L_1(\theta)=\frac{1}{M}\sum_{m=1}^{M}\left\|O^m-X^m\right\|_1 \tag{1}$$

Where O^m and X^m are the m-th reconstructed HR-HSI and ground truth, respectively. M is the number of images in a training batch, and θ denotes the parameter set of the network.

To simultaneously ensure that the reconstruction results have less spectral distortion, a Spatial Spectral Total Variation (SSTV) loss in [9] is introduced. It takes into account both spatial and spectral correlations.

$$L_{SSTV}(\theta)=\frac{1}{M}\sum_{m=1}^{M}\left(\left\|\nabla_h O^m\right\|_1+\left\|\nabla_w O^m\right\|_1+\left\|\nabla_l O^m\right\|_1\right) \tag{2}$$

Where ∇_h, ∇_w, and ∇_l denote the gradient functions of the computed horizontal, vertical, and spectral dimensions.

The final loss function can be expressed as follows:

$$L(\theta)=L_1(\theta)+\alpha L_{SSTV}(\theta) \tag{3}$$

Where α is the trade-off parameter, used to adjust the weight between space and spectral reconstruction error.

3. EXPERIMENTS AND RESULTS

Experiments are conducted on two public hyperspectral datasets, the CAVE dataset and the Harvard dataset. The CAVE dataset contains 32 scenes, with spectral bands (31 bands) acquired at intervals of 10nm in the range of 400-700nm. The Harvard dataset contains 77 HSIs of indoor and outdoor scenes. It also has 31 bands, and the spectral coverage range is 420-720nm. The spatial size of each image is 1040×1392.

Training and data simulation: In this paper, 20 images are selected from the CAVE dataset for training the network. The remaining 11 images are used for testing together with 9 images randomly selected from the Harvard dataset. Due to the small number of training samples available, this paper divides the training set images to obtain 4275 overlapping image patches with a size of 64×64×31 and regards them as ground truth. By applying a Gaussian filter with a blur kernel size of 3×3 and a standard deviation of 0.5, the overlapping image patches are down-sampled to a size of 16×16×31, which is regarded as the LR-HSI. In addition, spectral response functions of the Nikon D700 camera are used to generate HR-MSI patches. To train the proposed network, the Adam optimizer is applied with β_1=0.9, β_2=0.999, and the learning rate is initially equal to 1e−4, which is reduced by half every 75 epochs.

Performance Evaluation: In order to evaluate the performance, the proposed STSR method is compared with several state-of-the-art methods, including CNMF, FUSE, and Fusformer. Four quantitative evaluation indicators are used to measure the performance of different methods quantitatively, including peak signal-to-noise ratio (PSNR), spectral angle mapping (SAM), structural similarity (SSIM), and erreur relative global adimensionnelle de synthèse (ERGAS). The larger PSNR and SSIM are, and the smaller SAM and ERGAS are, the better the reconstruction effect is. Table 1 lists the reconstruction results of different methods on the two datasets when the scale factor is 4, where the optimal value is marked in bold, and the suboptimal value is marked with underline. It can be seen from Table 1 that the method proposed in this paper has reached the optimal or suboptimal value in each index. Figure 3 shows the results

obtained by different methods in a visual way. For easier observation and comparison, the details in the red box in the figure are enlarged.

Table 1. Comparative results of different methods with scale factor 4.

Dataset	Metrics	CNMF	FUSE	Fusformer	STSR
CAVE	PSNR	41.82	39.68	48.56	**49.97**
	SAM	7.32	4.97	2.52	**1.96**
	SSIM	0.975	0.979	0.995	**0.996**
	ERGAS	3.27	3.88	**1.30**	1.46
Harvard	PSNR	45.12	42.70	44.42	**45.98**
	SAM	2.56	2.69	2.66	**2.20**
	SSIM	0.978	0.970	**0.984**	0.980
	ERGAS	**2.05**	2.538	2.48	2.17

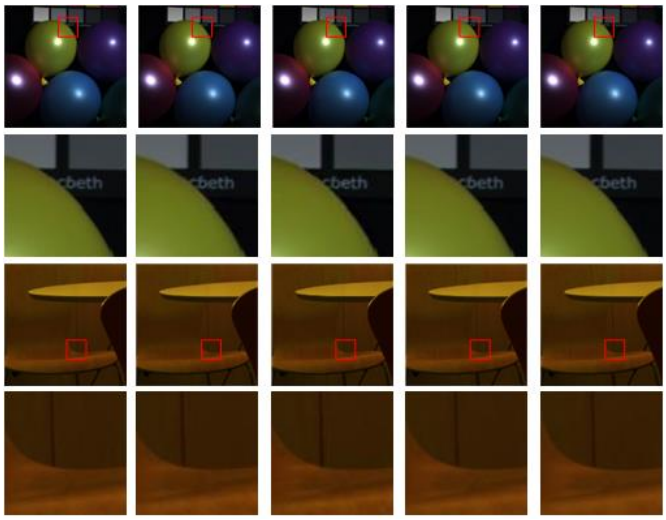

Figure 3. Visual results using different reconstruction methods with scale factor 4 for the CAVE and Harvard datasets. Spectral bands 26-14-6 are displayed as R-G-B display composite color images. To observe more clearly, each result's part of the red box is enlarged and shown in the 2nd and 4th row, respectively.

4. CONCLUSION

This paper proposes a hyperspectral image super-resolution algorithm based on Swin Transformer to fully extract spatial information and spectral similarity. Since CNN cannot effectively learn the global information of images, this paper leverages the Swin Transformer residual block to mine the contextual information of images with low computational cost using pixel-by-pixel input. Meanwhile, the spectral distortion of the reconstructed image is reduced by a spectral attention module and SSTV loss. Experimental results on two hyperspectral databases show that the proposed method can achieve better results and outperform the state-of-the-art methods. In addition, it is worth noting that it is also crucial to obtain multi-scale features of the image during the reconstruction process. The method proposed in this paper can achieve multi-scale feature extraction if different displacement window sizes are set at the Swin transformer residual block.

5. ACKNOWLEDGEMENTS

This work is supported by the National Nature Science Foundation of China (61801075, 42101350, 42271355), the Natural Science Foundation of Liaoning Province (2022-MS-160), the China Postdoctoral Science Foundation (No. 2020M670723), and the Fundamental Research Funds for the Central Universities (3132023238).

REFERENCES

[1] Camps-Valls G, Tuia D, Bruzzone L, et al. Advances in Hyperspectral Image Classification: Earth Monitoring with Statistical Learning Methods[J]. *IEEE Signal Processing Magazine*, 2013, 31(1): 45-54.

[2] Dian R, Li S, Sun B, et al. Recent Advances and New Guidelines on Hyperspectral and Multispectral Image Fusion[J]. *Information Fusion*, 2021, 69: 40-51.

[3] Li S, Dian R, Fang L, et al. Fusing Hyperspectral and Multispectral Images via Coupled Sparse Tensor Factorization[J]. *IEEE Transactions on Image Processing*, 2018, 27(8): 4118-4130.

[4] Xu T, Huang T Z, Deng L J, et al. Hyperspectral Image Super-resolution Using Unidirectional Total Variation with Tucker Decomposition[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2020, 13: 4381-4398.

[5] Dian R, Li S, Fang L. Learning A Low Tensor-train Rank Representation for Hyperspectral Image Super-resolution[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2019, 30(9): 2672-2683.

[6] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[J]. *arXiv preprint arXiv:2010*.11929, 2020.

[7] Liu Z, Lin Y, Cao Y, et al. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows[C]. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021: 10012-10022.

[8] Liang J, Cao J, Sun G, et al. Swinir: Image Restoration Using Swin Transformer[C]. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021: 1833-1844.

[9] Aggarwal H K, Majumdar A. Hyperspectral Image Denoising Using Spatio-spectral Total Variation[J]. *IEEE Geoscience and Remote Sensing Letters*, 2016, 13(3): 442-446.