SPECTRAL GROUPING DRIVEN HYPERSPECTRAL SUPER-RESOLUTION

Sadia Hussain, Brejesh Lall

BSTTM, IIT Delhi. Department of EE, IIT Delhi

ABSTRACT

Convolutional neural networks have proven to be proficient when extracting low-level concepts in an image. With the wonderful performance of transformers in exploiting the long-range correlations in an image, many methods have been explored where one exploit benefits of both the architectures. Therefore, in order to strengthen our network we add an important feature to transformers wherein single image super-resolution (SISR) is exploited using band grouping leveraging a simple CNN architecture. This paper aims to train a set of simple residual modelling architectures and then integrate them into a transformer architecture to solve super-resolution problem in HSI. We take a step forward to analyse how to adapt swinIR to fully exploit the information derived from band grouping for efficient SISR.

Index Terms— hyperspectral restoration, image super-resolution, transformers, convolutional neural networks, spatial super-resolution

1. INTRODUCTION

Hyperspectral images contain richer image information than RGB images which is typically how humans have perceived image data for years. This field has received a lot of attention with the increasing accuracy of hyperspectral sensors. Hence proving useful in many data processing techniques like image classification, image recognition, food security, anomaly detection, biomedical image processing and so on. The information is captured along spatial domain (2-D) and (1-D) spectral domain. However, capturing multi/hyper-resolution data is not without its challenges, and even after these challenges are addressed, one obtains data that is low resolution in the spatial dimension, or the spectral dimension, or both.

However, due to the limitation of imaging systems HSI resolution suffers even more. The first limitation is that the imaging hardware is not capable of simultaneous high spatial resolution and high spectral resolution. Therefore, these imaging systems can either acquire images with high spatial resolution (HR) with fewer spectral bands or images with high spectral resolution with poor quality spatial resolution (LR). This limitation has led to the reconstruction of high-resolution images (HSI) being considered as a super-resolution (SR) problem. Thus, super-resolution (SR) reconstructs good quality HSIs by increasing the resolution of an LR image to produce an HR image in a SR reconstruction technique. Broadly speaking, two types of spatial SR are explored: Fusion-based and single-image super-resolution (SISR). In the first category, multi-image super-resolution is formed using two sets of input vis-a-vis LR HSI and an auxiliary image HR. This auxiliary image can be any image with high spatial superresolution, such as an MSI (multispectral image) or a panchromatic image (PAN). SISR, on the other hand, does not use an auxiliary image. Methods of super-resolving a single image include interpolation, tensor-based methods, Bayesian methods, and deep learning approaches. Since the absolute goal is spatial super-resolution, we give the low-resolution HSI (such as LR-HSI) as input to a deep learning based approach (DL). As a learning-based method, DL has recently used transformers for image recovery. Given the insufficient training dataset in HSI, transformer-based methods are largely suitable for RGB tasks. Moreover, the data in HSI can be expressed in two ways: spatial and spectral. Transformers generally exploit spatial correlations better, and it is challenging to exercise control over the spectral domain. As we will be see in this work the shallow encoder overcomes this shortcoming using a residual learning architecture.

To overcome the above mentioned challenges, we use a shallow encoder-decoder architecture with transformer, which is used to extract deep features. The shallow encoder architecture is a simple yet effective architecture that allows for better grouping as well as residual modelling as a learnable parameter. Grouped bands help realise effective spatial SR performance based on spectral reflectance. Bands grouped in this way better contribute to creating an effective pixel-wise feature map. In addition to grouping the bands, the residual modelling architecture in the shallow encoder allows adaptive rescaling between interdependent channels within each band group exploit spectral dependencies. This rich embedding is then fed into a transformer module. To use these shallow and deep features, we use a special upscale filter called sub-pixel convolution as our shallow decoder. We summarize the main contributions of the paper as follows:

- We propose a spectral grouping mechanism that exploits the correlation structure of the Hyperspectral bands, resulting in better features. The band groups are processed by shallow residual network consisting of a

Thanks to XYZ agency for funding.

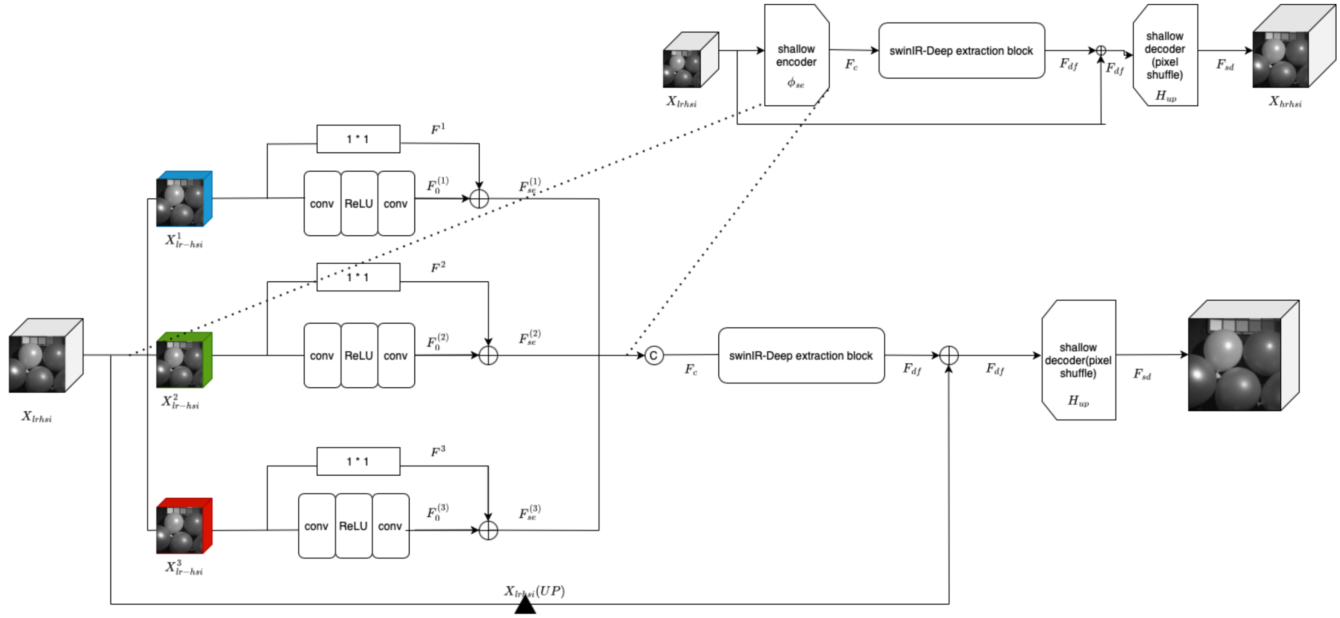


Fig. 1. Spatial Grouping driven Hyperspectral Superresolution architecture. The proposed shallow encoder is expanded to illustrate its structure.

cascade of Convolution, ReLU and a second Convolution.

- The shallow features are processed using the SwinIR architecture to generate high quality deep high-level features.
- These high level features are processed using a shallow detector comprising of the pixel shuffle architecture, adapted to handle multiple bands instead of just 3 in case of RGB images.
- An optional global residual connection is added to further enhance the quality of the superresolved hyperspectral image.

1.1. Related Works

To address the problem of super-resolution in HSI, two categories, fusion-based super-resolution and single-image super-resolution are distinguished HSI-SR techniques. Initially, Deep Learning (DL) based architectures were quite widely used and have shown good performance due to their exceptional feature representation. Many studies consider DL, especially the CNN method, for fusing LR-HSI and HR-MSI, which have outperformed many conventional methods [1, 2, 3, 4, 5]. However, there are still opportunities for improvement since convolution has a limited capacity for information extraction. Vaswani et al [6] originally proposed transformers for natural language processing (NLP). With their large representation capacity, transformers enable remarkable advances. Transformer has been extended to a variety of computer vision tasks in recent years and has emerged as a strong competitor to CNN in vision applications such as image recognition [7], segmentation [2], and object recognition [8]. In addition, transformer has been developed to handle low-level vision tasks such as image restoration [3, 9, 2]. The SwinIR model for image restoration based on the Swin transformer was proposed by [7] to apply self-awareness in local space. Transformers are now finding applications in HSI as well. Transformers with their variants such as vision transformers have been proposed to utilise spectral awareness and sometimes spatial/spectral features simultaneously. Recently, the fusformer [10] was the first of its kind to be used in HSI to obtain HR-HSI from HR-MSI and LR-HSI, using a fusion technique. Also [11] has exploited the use of 3D ViT for HSI classification tasks. Based on token embedding, [12] uses patchwise learning to exploit semantic tokenisation for HSI classification task. Recently [13] uses transformers to exploit spectral and spatial dependencies. However, this method uses RGB restoration as an auxiliary task. This work is thus the first of its kind to perform transformer based HSI restoration as SISR.

The rest of the paper is organised as follows: Section 2 contains a description of the proposed architecture. Section 3 is the description of the experiments performed, performance measures used, and an analysis of the results obtained. Finally, section 4 contains the conclusions.

2. METHODOLOGY

2.1. Overall Network Architecture

As shown in **Figure 1**, our network architecture consists of three parts: a shallow encoder block, followed by a deep high-level feature extraction block, and a reconstruction block, which is the shallow decoder. The shallow encoder block operates independently on three groups of bands. Each branch consists of a cascaded of 3×3 convolutional layer, ReLU, and a second 3×3 convolutional layer, along with a skip connection consisting of a 1×1 residual learning block. Shallow encoder ensures better correlation learning for each group. With the feature maps obtained in this layer, a deep feature extraction layer is used to further extract the lost high frequencies using a transformer architecture. For this purpose we adapt swinIR which is based on shifted window attention. Finally, we use the subpixel convolutional layer as a reconstruction module to exploit the features of the shallow and deep layers.

2.2. Initial restoration block and Deep feature extraction:

The spectral response function (SRF) can effectively provide better spatial superresolution across the of set of low to high resolution bands. SRF can be used as a guide to constrain spatial information to achieve better visual quality of images, more details, higher resolution, and better edge detection. At the same time, it's also used to constrain spectral information such as high correlations between adjacent bands and high cross-correlations between distant bands. In this paper, a shallow encoder block is proposed to group bands by spectral radiation characteristics. These bands are divided by red, green, and blue, and the number of bands in each group depends on the amount of scattered spectral radiation. We find that this concept has been explored in [14] spectral superresolution, but has not yet been used in spatial superresolution. We remedy that here, and do so in this work. Given a low-resolution input signal $X_{lrhsi} \in R^{h \times w \times C}$ (where h, w, C stand for height, width and channel, downsampled by $2\times$) we split the spectral ranges in three groups $X_{lr-hsi} = \{X^1_{lr-hsi}, X^2_{lr-hsi}, X^3_{lr-hsi}\}$ according to the amount of spectral correlation a group has to offer. Once the groups are formed, they are fed into the residual modelling architecture ϕ_{se}. Each ϕ_{se} block has two parts: a spatial module and a residual learning module as shown in **Figure 1**. For the spatial module, we use a simple 3×3 Conv-ReLU-Conv block to extract the spatial features. However, since a standard convolution operation isn't able to extract the spectral dependencies between spectra in a hyperspectral image, we employ a residual channel attention using a 1×1 convolution. Within each residual module, $K_{GC}(.)$ convolution is applied. This convolution extracts the low-level features $F_0 \in R^{h \times w \times d}$ for each group $X^{(s)}_{lrhsi}$ as:

$$F_0^{(s)} = K_{GC}(X^{(s)}_{lrhsi}) \tag{1}$$

where $F_0^{(s)}$ refers to the initial feature map of the shallow encoder. To this, a residual learning parameter module $\psi(.)$ is added. $\psi(.)$ is a 1×1 convolution acting on one of the input groups such that:

$$F^i = \psi(X^{(s)}_{lrhsi}) \tag{2}$$

Here F^i is the feature map pertaining to residual learning module. By constructing a spectral network with $\psi(.)$ filters, we can assume that the correlations between the different spectral bands are fully exploited. Finally the element-wise residual output of the s^{th} group branch with a residual learning is given by $F_{se}^{(s)}$ as:

$$F_{se}^{(s)} = F_0^{(s)} + \psi(X^{(s)}_{lrhsi}), \quad s = 1, 2, 3 \tag{3}$$

$$F_{se}^{(s)} = F_0^{(s)} + F^s, \quad s = 1, 2, 3 \tag{4}$$

After extracting the branch outputs the features are concatenated. The concatenated features denoted as $F_c = [F^1, F^2, F^3]$ are then input to swinIR. With impressive results in RGB, we use a swinIR, composed of K swin transformers assembled in a residual form with a convolutional layer at the end. This in turn consists of multiple swin transformer layers, such that:

$$F_{df} = H_{swinIR}(F_c) \tag{5}$$

2.3. Final restoration block

The output of the deep feature extraction layer have spatial-spectral features and can then be fed into a reconstruction layer to produce the final high-resolution hyperspectral images. This reconstruction is called an upscaling module. This upscaled module takes the aggregated features and, using the advantages of subpixel convolution (pixel shuffle), carries forward the enhancement. Three 3×3 convolutional layers and the residual connection are part of this residual enhancement module [15].

$$F_{sd} = H_{up}(F_{df}) \tag{6}$$

where $H_{up}(.)$ is the upsampled sub-pixel convolution over the output of the transformer block F_{df}

2.4. Loss function

To evaluate the reconstruction accuracy of the network in this case, we use the L1 loss, since it can efficiently penalise errors and ensure better convergence during the training phase. The mean absolute error (MAE) between all reconstructed images and the ground truth is what specifically defines the L1 loss.

$$\mathcal{L}oss = \frac{1}{N} \sum_{n=1}^{N} |X^n_{HR} - H_{net}(X^{(n)}_{lrhsi})| \tag{7}$$

NTIRE2022				
Methods	RMSE	PSNR	SAM	SSIM
Ours	0.0107	39.37	1.53	0.9823

Table 1. Results conducted on NTIRE 2022 dataset with our proposed method

where N is the number of images in the training dataset and X^n_{HR}, $H_{net}(X^{(n)}_{lrhsi})$ are n^{th} ground truth and reconstructed HSI respectively.

3. EXPERIMENTS

3.1. Brief description of datasets and experimental design

In this work, two datasets are used to evaluate the proposed method: NTIRE2022 [16] and Cave [17] dataset. NTIRE2022 is a spectral recovery challenge dataset with $512 \times 438 \times 31$ sized dataset. The number of bands in NTIRE2022 are 31 bands which range from 400nm to 700nm. On the other hand, the Cave dataset is $512 \times 512 \times 31$ with a spectral range from 430nm to 730nm.

We use Adam's optimizer with a learning rate set to 2e-4. The batch size is 32. The number of swin transformer blocks is set to 6 with a depth of 6 and a window size of 8. Number of attention heads is reserved as 6. We augment the size of training dataset to 512×512 and then downsample by $X2$. The feature embedding size of the training dataset is set to 96 in the proposed approach. The experiments are performed using Pytorch framework. The performance of super-resolution performance is evaluated using four performance metrics such as PSNR, RMSE, SSIM and SAM.

3.2. Experiment results

The quantitative results are given in Table 1 and Table 2. Our method improves PSNR and other metrics on two datasets. We compare the performance of our proposed method with other current existing methods such as SSPSR [18], DeepPrior [19], HSISR [5], DSTrans [13]. However, SSPSR [18] is a regular DCNN based method, however, Fusformer [10] is a transformer based method. HSISR [5], on the other hand is a encoder-decoder based architecture. HSISR [5] and DSTrans [13] use input as auxiliary RGB learning. Fusformer [10], on the other hand, is a fusion based method. SSPSR [18] is the only method that uses a similar flow and therefore provides the best indication of the strength of our proposed method.

We report our main findings on the NTIRE2022 dataset. Since this dataset is very new, the above methods did not perform their results on this dataset. Therefore, the common dataset selected for comparison of our method with other existing methods is Cave. The results for the Cave dataset with other existing methods are shown in Table 2

CAVE				
Methods	RMSE	PSNR	SAM	SSIM
DeepPrior [19]	0.0141	37.060	3.410	0.9418
HSISR [5]	0.0131	39.060	3.360	0.9618
SSPSR [18]	0.0136	38.302	3.36	0.9566
DSTrans [13]	0.0118	40.073	3.169	0.9659
Ours	**0.0085**	**41.37**	**1.70**	**0.9906**

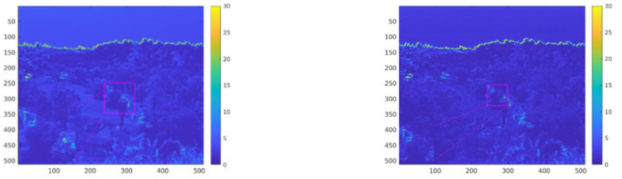
Table 2. Results conducted on CAVE dataset with our proposed method

3.3. Ablation Study

The ablation study results are reported for the NTIRE2022 and cave datasets. We perform experiments related to the use and non-use of residual modelling in our architecture. The 'w/o' means that residual modelling was not used, and 'w' means that residual modelling was used.

w/o residual modelling				
Datasets	RMSE	PSNR	SAM	SSIM
CAVE	0.00951	40.47	1.71	0.981
NTIRE 2022	0.0150	36.486	4.756	0.9592

w residual modelling				
Datasets	RMSE	PSNR	SAM	SSIM
CAVE	0.0085	41.37	1.70	0.9906
NTIRE 2022	0.0107	39.37	1.53	0.9823



(a) w/o residual modelling

(b) w residual modelling

Fig. 2. Here (a) represents no residual modelling whereas (b) represents residual modelling architecture. The above figure show the error maps for NTIRE2022 dataset

4. CONCLUSION

In this study, a unified spectral-spatial channel grouping based swin transformer adaptation is proposed. Our model explores the connection between the spectral correlation and spatial correlation information from the grouping of channels based on the spectral reflectivity. The product of channel grouping using a shifted window based transformer, results in a richer embedding. This can be exploited to achieve better visual quality performance. The upsampling method used solves the problem of limited number of training samples in HSI. The evaluation of three datasets using our proposed method leads to generate high-quality super-resolution HSI.

5. REFERENCES

[1] Yunsong Li, Jing Hu, Xi Zhao, Weiying Xie, and JiaoJiao Li, "Hyperspectral image super-resolution using deep convolutional neural network," *Neurocomputing*, vol. 266, pp. 29–41, 2017.

[2] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12299–12310.

[3] Zhengyu Liang, Yingqian Wang, Longguang Wang, Jungang Yang, and Shilin Zhou, "Light field image super-resolution with transformers," *IEEE Signal Processing Letters*, vol. 29, pp. 563–567, 2022.

[4] Xiaolin Han, Jing Yu, Jiqiang Luo, and Weidong Sun, "Hyperspectral and multispectral image fusion using cluster-based multi-branch bp neural networks," *Remote Sensing*, vol. 11, no. 10, pp. 1173, 2019.

[5] Ke Li, Dengxin Dai, and Luc Van Gool, "Hyperspectral image super-resolution with rgb image super-resolution as an auxiliary task," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 3193–3202.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[7] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1833–1844.

[8] Z Wang, X Cun, J Bao, W Zhou, J Liu, and H Uformer Li, "A general u-shaped transformer for image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA*, 2022, pp. 19–24.

[9] Haobo Ji, Xin Feng, Wenjie Pei, Jinxing Li, and Guangming Lu, "U2-former: A nested u-shaped transformer for image restoration," *arXiv preprint arXiv:2112.02279*, 2021.

[10] Jin-Fan Hu, Ting-Zhu Huang, Liang-Jian Deng, Hong-Xia Dou, Danfeng Hong, and Gemine Vivone, "Fusformer: A transformer-based fusion network for hyperspectral image super-resolution," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[11] Weilian Zhou, Sei-Ichiro Kamata, Zhengbo Luo, and Xi Xue, "Rethinking unified spectral-spatial-based hyperspectral image classification under 3d configuration of vision transformer," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 711–715.

[12] Danfeng Hong, Zhu Han, Jing Yao, Lianru Gao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot, "Spectralformer: Rethinking hyperspectral image classification with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2021.

[13] Dabing Yu, Qingwu Li, Xiaolin Wang, Zhiliang Zhang, Yixi Qian, and Chang Xu, "Dstrans: Dual-stream transformer for hyperspectral image restoration," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3739–3749.

[14] Jiang He, Jie Li, Qiangqiang Yuan, Huanfeng Shen, and Liangpei Zhang, "Spectral response function-guided deep optimization-driven network for spectral super-resolution," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 9, pp. 4213–4227, 2021.

[15] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.

[16] Boaz Arad, Radu Timofte, Rony Yahel, Nimrod Morag, Amir Bernat, Yaqi Wu, Xun Wu, Zhihao Fan, Chenjie Xia, Feng Zhang, et al., "Ntire 2022 spectral demosaicing challenge and data set," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 882–896.

[17] F. Yasuma, T. Mitsunaga, D. Iso, and S.K. Nayar, "Generalized Assorted Pixel Camera: Post-Capture Control of Resolution, Dynamic Range and Spectrum," Tech. Rep., Nov 2008.

[18] Junjun Jiang, He Sun, Xianming Liu, and Jiayi Ma, "Learning spatial-spectral prior for super-resolution of hyperspectral imagery," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1082–1096, 2020.

[19] Oleksii Sidorov and Jon Yngve Hardeberg, "Deep hyperspectral prior: Single-image denoising, inpainting, super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.