3D Expansion of SRCNN for Spatial Enhancement of Hyperspectral Remote Sensing Images

Nour Aburaed*, Mohammed Q. Alkhatib†, Stephen Marshall‡, Jaime Zabalza§, Hussain Al Ahmad¶
*†¶ College of Engineering and IT, University of Dubai, UAE
*‡§ Department of Electronic and Electrical Engineering, University of Strathclyde, UK
Email: *nour.aburaed,†mqalkhatib@ieee.org
*nour.aburaed, ‡j.zabalza, §stephen.marshall@strath.ac.uk, ¶halahmad@ud.ac.ae

***Abstract*—Hyperspectral Imagery (HSI) have high spectral resolution but suffer from low spatial resolution due to sensor tradeoffs. This limitation hinders utilizing the full potential of HSI. Single Image Super Resolution (SISR) techniques can be used to enhance the spatial resolution of HSI. Since these techniques rely on estimating missing information from one Low Resolution (LR) HSI, they are considered ill-posed. Furthermore, most spatial enhancement techniques cause spectral distortions in the estimated High Resolution (HR) HSI. This paper deals with the extension and modification of Convolutional Neural Networks (CNNs) to enhance HSI while preserving their spectral fidelity. The proposed method is tested, evaluated, and compared against other methodologies quantitatively using Peak Signal-to-noise Ratio (PSNR), Structural Similarity Index Measurement (SSIM), and Spectral Angle Mapper (SAM).**

***Index Terms*—Hyperspectral, remote sensing, single image super resolution, 3D convolution**

I. Introduction

The aggregate and steady progress of remote sensing imagery since the 1950s allowed for rapid advancements in several industrial facets, such as agriculture, surveillance, and urban development. In conjunction with the fields of image processing and artificial intelligence, remote sensing efficiency is boosted through automating key tasks, such as object detection, semantic segmentation, and classification, which saves time and effort due to minimizing human intervention. However, automating such tasks with high accuracy is not an easy feat. Nowadays, remote sensing technology facilitates various image resolutions that render image processing and automation tasks relatively easier. The challenge relies in the inherent trade-off of sensors, which makes them capable of capturing either images with high spectral resolution, or high spatial resolution. Hyperspectral Images (HSIs) consist of hundreds of contiguous bands that cover a wide range of wavelengths. Essentially, an HSI is a 3D data cube, where each pixel position across all bands can be represented with spectral reflectance values, known as spectral signature. HSIs have high spectral resolution, but suffer from low spatial resolution that hinders exploiting their full potential. Hence, improving the spatial resolution of HSI, also known as HSI Super Resolution (HSI-SR), is a highly important pre-processing step for HSI applications. Broadly, HSI-SR approaches can be categorized into fusion methods, and Single Image Super Resolution (SISR) methods. Fusion methods suffer from various shortcomings, such as spectral distortions and the requirement of auxiliary information. Therefore, SISR approaches have been widely used for the improvement of spatial resolution of HSI, especially after the success of SISR in RGB enhancement. SISR refers to the process of generating a High Resolution (HR) image from a single Low Resolution (LR) image. It is considered as an ill-posed problem due to the difficulty of estimating the missing high frequency components from the LR image without auxiliary information. Traditionally, SISR approaches started with interpolation methods, which include nearest neighbor, bilinear, and bicbuic interpolation. Interpolation methods suffer from blurriness and artifacts due to aliasing effects, which are especially evident with high scaling factors, and that renders them ineffective for HSIs. Nowadays, Convolutional Neural Networks (CNNs) and Deep Learning methods are considered as the most effective approaches for SISR. CNNs are a type of Artificial Neural Networks (ANNs) that are particularly effective for image processing tasks. When a CNN is built for SISR, it typically consists of a combination of three main layers: i) convolution, ii) activation function, iii) pooling. There are many examples in the literature of of SISR CNNs that enhance RGB or, more generally, multispectral images [1]–[4]. The most prominent ones include SRCNN [5], Very Deep SR (VDSR) [6], and SR Generative Adversarial Network (SRGAN) [7]. These networks cannot be directly used for enhancing HSIs; ignoring the main difference between HSIs and RGB images, which is the spectral resolution, causes spectral distortions in the spatially enhanced outcome. Therefore, spectral fidelity must be accounted for. The aforementioned networks perform calculations in 2D, i.e. only the spatial context is considered. However, if these networks were extended to 3D, then the spectral aspect can be accommodated as well, which is the main hypothesis of this paper. Many HSI-SR attempts in the literature support the effectiveness of 3D CNNs when it comes to processing HSIs [8]–[11]. In this paper, one of the most prominent CNNs for RGB enhancement, namely SRCNN, will be extended to 3D domain and adapted for HSI-SR. The network will be trained and evaluated using Pavia University dataset. The performance will be assessed qualitatively in terms of spectral signature, and quantitatively in terms of Peak Signal-to-Noise Ratio (PSNR),

Structural Similarity Index Measurement (SSIM), and Spectral Angle Mapper (SAM). The results of the 3D SRCNN will be compared to its original 2D version, as well as other state-of-the-art algorithms. Furthermore, the sizes of both 2D and 3D SRCNNs filters are modified and the performance is compared to the original filters The rest of the paper is organized as follows: Section II showcases the dataset used in this study, Section III explains the details of the methodology and training scheme, Section IV illustrates the results and draws comparisons and deductions, and finally, Section V summarizes and concludes the paper.

II. Dataset

Pavia University is an HSI dataset that consists of one scene acquired by ROSIS sensor. The scene has 103 bands of size 610×610 pixels. The spectral range covered by ROSIS sensor is 430-960nm, where each band is allocated 5nm. DCNNs require a large dataset size for training. Therefore, the scene is divided into blocks of 64×64 to fulfill training and testing requirements. The total number of HSIs acquired is 41 for training and validation and 4 for testing. For each block, a LR counterpart is generated synthetically via bicubic interpolation and Gaussian blur for training and testing DCNNs over scaling factors of 2 and 4.

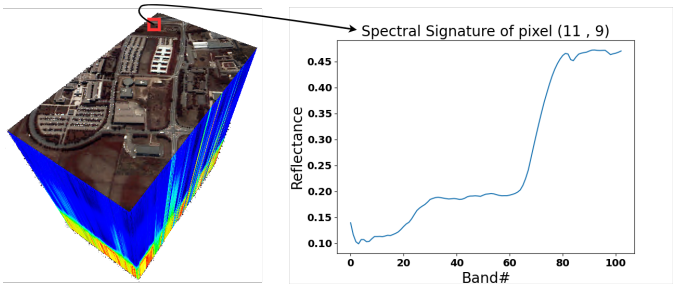
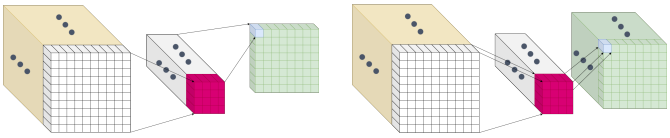


Fig. 1: Sample hyperspectral 3D cube and spectral signature of pixel location (11,9) from Pavia University.

III. Methodology

SRCNN is a 2D CNN that was developed for the purpose of enhancing the spatial resolution of RGB and/or gray images. The network operates on band-by-band basis due to its 2D nature. One of the advantages of the network is the fact that it is lightweight, as it consists of three 2D convolution layers only, each one followed by a Rectified Linear Activation Function (ReLu). A convolution layer is basically a filter that consists of several kernels, which contain input weights. In the case where the filter contains only one kernel, these two terms can be used interchangeably. The output of a 2D convolution is a feature map generated from an input image. For each input pixel, its feature value is computed by adding the product of its neighboring pixels and the corresponding weight in the kernel. For an image I of size $N \times N$ and a kernel K of size



(a) 2D convolution (b) 3D convolution

Fig. 2: Visual illustration of 2D and 3D convolution.

$M \times M$, 2D convolution at position (x, y) can be expressed as the following equation:

$$Y_{(x,y)} = ReLU\left(\sum_i \sum_j K_{(i,j)} I_{(x+i,y+j)} + b\right), \quad (1)$$

where $Y_{(x,y)}$ is the output feature, $I_{(x+i,y+j)}$ is the input pixel that includes the original pixel and the neighboring pixels within the offset range (i, j), $K_{(i,j)}$ is the weight at location (i, j) that corresponds to the input, and b is the bias. Convolution process causes the input to lose spatial dimensionality, such that the size of the output feature map is smaller than the size of the input. This can be avoided by padding the input at the borders, which will force the output size to match the original input. The simplest padding method is performed by adding zeroes at the borders of the input.
2D convolution works well for multispectral images. However, when considering an HSI cube with hundreds of bands, processing each band individually may lead to acceptable spatial enhancement, but it causes spectral distortions due to ignoring spectral context and correlation between bands. This means an HSI must be processed as a cube and, therefore, 3D convolution is an adequate solution to accommodate spectral context. Figure 2 provides a visual comparison between 2D and 3D convolutions. Rather than processing each band individually, 3D convolution performs computations over the height, width, and bands. For an image I of size $N \times N \times B$ and a kernel K of size $M \times M \times F$, 3D convolution at position (x, y, z) can be expressed with the following equation:

$$Y_{(x,y,z)} = ReLU\left(\sum_i \sum_j \sum_k K_{(i,j)} I_{(x+i,y+j,z+k)} + b\right) \quad (2)$$

Based on this convention, SRCNN can be extended to 3D by converting all of its 2D convolution layers to 3D ones. The overall architecture of the 3D SRCNN can be seen in Figure 3. SRCNN originally has filters of sizes (9,9), (1,1), and (5,5). The extended 3D version has filters of sizes (9,9,9), (1,1,1), and (5,5,5). The images are padded in such a way that their sizes is not reduced. The padding is calculated as follows:

$$\begin{aligned} N + (2 \times padding) - (M - 1) &= N \\ Padding &= \frac{M-1}{2} \end{aligned} \quad (3)$$

In this case, the input image is padded with 5 zeros on all sides, which negatively affects the overall quality of the output. Therefore, a modified architecture of both SRCNNs

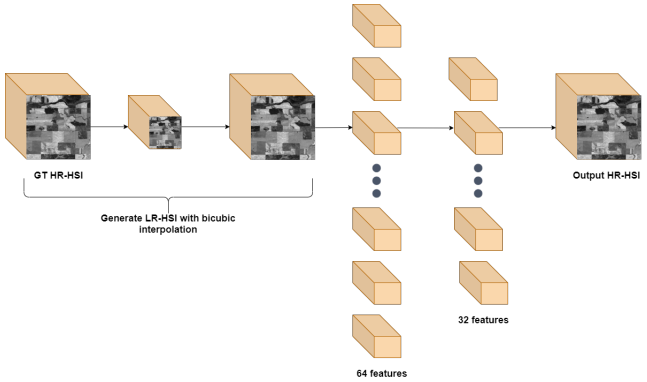


Fig. 3: Overall Architecture of SRCNN 3D.

was created to reduce padding effects and comapre their performance against the original SRCNNs. The new filter sizes are all reduced to (3,3) in case of 2D, and (3,3,3) in case of 3D. A comparison between filter sizes of all layers is provided in Table I.

TABLE I: Comparison between the original and modified SRCNN 2D and 3D architectures.

Layer #	SRCNN 2D		SRCNN 3D	
	Original filter	Modified filter	Original filter	Modified filter
1	(9,9)	(3,3)	(9,9,9)	(3,3,3)
2	(1,1)	(3,3)	(1,1,1)	(3,3,3)
3	(5,5)	(3,3)	(5,5,5)	(3,3,3)

Both 2D and 3D networks were trained under the same conditions. The training parameters can be seen in Table II. These parameters were set empirically based on the best performance obtained. The learning rate value is not fixed, as it is adjusted each epoch according to the performance of the network. Since training data are shuffled, each network is trained and tested 10 times, and then the average results of all 10 experiments are recorded. The next section discusses the results and comparisons.

TABLE II: Parameters used to train 2D and 3D SRCNN.

Parameter	**Value**
Epochs	250
Learning rate	1×10^{-5}
Optimization	Adam
Loss	MSE
Batch size	2
Shuffle	True

IV. Results and Discussion

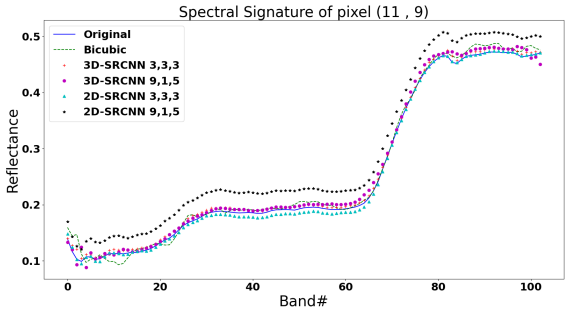
The training was done using Tensorflow-GPU library on NVIDIA Quadro P6000-24GB X 2 GPU with 380GB RAM. Quantitative evaluation is the best way to compare between the predicted output and the groundtruth (GT) objectively. One of the most commonly used quantitative metrics is PSNR, which is expressed as follows:

$$PSNR = 10log_{10}\frac{MAX(HS)^2}{MSE},$$
$$MSE = \frac{1}{M \times N}\sum_{i=1}^{N}\sum_{j=1}^{M}[HS(i,j) - \widetilde{HS}(i,j)]^2, \quad (4)$$

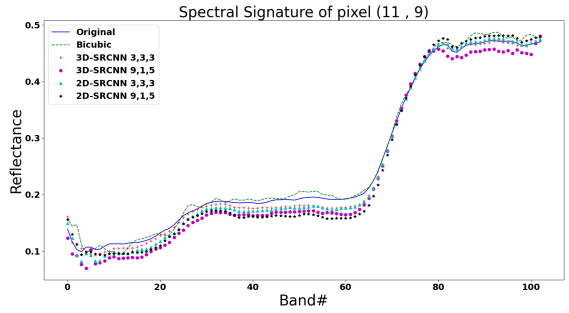
where HS is one GT HSI band and $\widetilde{HS}$ is the estimated HSI band. $MAX(HS)$ is the maximum possible value a single pixel band can take. For this study, all images have been converted to unsigned integer 8, so the maximum value is 255. If $\widetilde{HS}$ is identical to HS, PSNR value would be infinite. Since this outcome does not occur in practice, PSNR value should be as high as possible. One of the shortcomings of PSNR is that it does not take the human visual perception into consideration. Therefore, it is not a reliable metric by itself. SSIM is a good measurement of human visual perception because it assesses the error of three components; correlation, contrast, and illumination, which are expressed in equation 5.

$$SSIM = \frac{(2\mu_{HS}\mu_{\widetilde{HS}} + C_1)(2\sigma_{HS\widetilde{HS}} + C_2)}{(\mu_{HS}^2 + \mu_{\widetilde{HS}}^2 + C_1)(\sigma_{HS}^2 + \sigma_{\widetilde{HS}}^2 + C_1)}, \quad (5)$$

where μ_{HS}, $\mu_{\widetilde{HS}}$, σ_{HS}, $\sigma_{\widetilde{HS}}$, and $\sigma_{HS\widetilde{HS}}$ represent local means, standard deviation and cross-covariance for images HS and $\widetilde{HS}$. SSIM value ranges between 0 if the images being



(a) $\times 2$



(b) $\times 4$

Fig. 4: Plots of reflectance signatures of pixel (11,9) produced by each algorithm in addition to the GT.

compared are poles apart, and 1 if the images are identical. In practice, SSIM should be as close to 1 as possible.

Both PSNR and SSIM do not capture spectral context. SAM is a useful quantitative metric that covers this missing aspect, which determines the similarity between the spectra of the GT HSI and the spectra of the enhanced HSI. SAM is expressed as follows:

$$SAM = cos^{-1}\left(\frac{\sum_{i=1}^{B} HS_i \widetilde{HS}_i}{\sqrt{\sum_{i=1}^{B} HS^2}\sqrt{\sum_{i=1}^{B} \widetilde{HS}^2}}\right). \quad (6)$$

SAM value is measured in degrees, and it should be as close to 0 as possible. The aforementioned metrics were used to evaluate the quality of the enhanced images from all SRCNN 2D and 3D networks. The results are summarized in Table III along with the enhanced HSIs using bilinear and bicubic interpolation. It can be observed that all versions of SRCNN 2D and 3D prevail over bicubic and bilinear interpolation, with the exception of SRCNN 2D(9,1,5) in scale factor 2. The (3,3,3) version of SRCNN 2D prevails over (9,1,5) version in both factors 2 and 4. A similar observation can be made about SRCNN 3D. For (9,1,5), SRCNN 3D shows considerable enhancement over SRCNN 2D. However, for (3,3,3), the enhancement is greater by only a small margin for scale factor 2, and the PSNR for SRCNN 3D is slightly less than SRCNN 2D for scale factor 4. Overall, SRCNN 3D (3,3,3) shows the best performance in terms of PSNR, SSIM, and SAM. Figure 4 shows the spectral signature of pixel (11,9) chosen at random. the spectral signature is plotted for every result predicted by 2D and 3D network, in addition to the result obtained from bicubic interpolation. The spectral signatures are consistent with the results listed in Table III, as SRCNN 2D (9,1,5) shows the most spectral distortions, while SRCNN 3D (3,3,3) shows the least spectral distortions compared to others. The spectral signature of SRCNN 2D (3,3,3) also shows a pattern similar to GT, especially for scale factor 4 seen in Figure 4b.

TABLE III: Results summary of all SRCNN 2D and SRCNN 3D networks im comaprison with bilinear and bicubic interpolation for factors of $\times 2$ and $\times 4$.

Scale Factor	Method	PSNR (dB)	SSIM	SAM (°)
$\times 2$	SRCNN 2D(9,1,5)	26.45	0.76	10.20
	SRCNN 2D(3,3,3)	31.12	0.90	5.82
	SRCNN 3D(9,1,5)	29.37	0.87	8.62
	SRCNN 3D(3,3,3)	**31.92**	**0.91**	**5.33**
	Bicubic	29.27	0.85	5.56
	Bilinear	28.61	0.81	5.90
$\times 4$	SRCNN 2D(9,1,5)	25.39	0.69	9.06
	SRCNN 2D(3,3,3)	**27.31**	0.75	**7.29**
	SRCNN 3D(9,1,5)	26.97	0.75	8.31
	SRCNN 3D(3,3,3)	27.19	**0.76**	7.63
	Bicubic	24.86	0.64	8.58
	Bilinear	24.70	0.62	8.72

V. Conclusion

In this paper, one of the most prominent networks for RGB enhancement, namely SRCNN, was modified and extended to 3D in order to enhance HSI spatially while avoiding spectral distortions. Four versions of the networks were developed, trained, and tested, including SRCNN 2D (9,1,5) and (3,3,3), and SRCNN 3D (9,1,5) and (3,3,3). Quantitative evaluation shows that both (3,3,3) versions of the network prevail over (9,1,5). Additionally, all networks aside from SRCNN 2D (9,1,5) prevail over the traditional bicubic interpolation. SRCNN 3D (3,3,3) shows the best performance for scale factor 2, while it sometimes falls slightly behind SRCNN 2D (3,3,3). Visualizing the spectral signature of each network's outcome shows consistent results to quantitative evaluation, as SRCNN 3D (3,3,3) and SRCNN 2D (3,3,3) appear the closest to the GT spectral signature. The next steps of this study involve further enhancements to SRCNN 3D (3,3,3) to further eliminate spectral distortions and boost the network's performance for higher scale factors.

References

[1] N. Aburaed, A. Panthakkan, S. Almansoori, and H. Al Ahmad, "Super resolution of DS-2 satellite imagery using deep convolutional neural network," in *Image and Signal Processing for Remote Sensing XXV*, Lorenzo Bruzzone and Francesca Bovolo, Eds. International Society for Optics and Photonics, 2019, vol. 11155, pp. 485–491, SPIE.

[2] N. Aburaed, A. Panthakkan, M. Al-Saad, M. C. El Rai, S. Al Mansoori, H. Al Ahmad, and Stephen Marshall, "Super-resolution of satellite imagery using a wavelet multiscale-based deep convolutional neural network model," in *Image and Signal Processing for Remote Sensing XXVI*, Lorenzo Bruzzone, Francesca Bovolo, and Emanuele Santi, Eds. International Society for Optics and Photonics, 2020, vol. 11533, pp. 305–311, SPIE.

[3] J. Yamanaka, S. Kuwashima, and T. Kurita, "Fast and accurate image super resolution by deep CNN with skip connection and network in network," *CoRR*, vol. abs/1707.05425, 2017.

[4] H. Huang, L. Shen, C. He, W. Dong, H. Huang, and G. Shi, "Lightweight image super-resolution with hierarchical and differentiable neural architecture search," 2021.

[5] C. Dong, C. L. Change, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham, 2014, pp. 184–199, Springer International Publishing.

[6] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1646–1654.

[7] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," *CoRR*, vol. abs/1609.04802, 2016.

[8] L. Wang, T. Bi, and Y. Shi, "A frequency-separated 3d-cnn for hyperspectral image super-resolution," *IEEE Access*, vol. 8, pp. 86367–86379, 2020.

[9] Q. Li, Q. Wang, and X. Li, "Mixed 2d/3d convolutional network for hyperspectral image super-resolution," *Remote Sensing*, vol. 12, no. 10, 2020.

[10] 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 Sensing*, vol. 9, no. 11, 2017.

[11] S. Mei, X. Yuan, J. Ji, S. Wan, J. Hou, and Q. Du, "Hyperspectral image super-resolution via convolutional neural network," in *IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 4297–4301.