SISR of Hyperspectral Remote Sensing Imagery Using 3D Encoder-Decoder RUNet Architecture

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

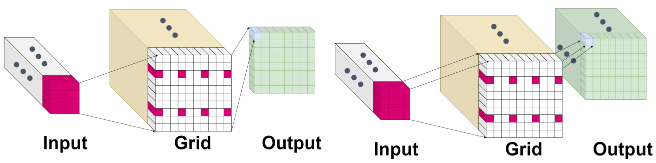
***Abstract*—Single Image Super Resolution (SISR) refers to the spatial enhancement of an image from a single Low Resolution (LR) observation. This topic is of particular interest to remote sensing community, especially in the area of Hyperspectral Imagery (HSI) due to their high spectral resolution but limited spatial resolution. Enhancing the spatial resolution of HSI is a pre-requisite that boosts the accuracy of other image processing tasks, such as object detection and classification. This paper deals with SISR of HSI through the 3D expansion of Robust UNet (RUNet). The network is developed, trained, and tested over two datasets, and compared against the original 2D-RUNet and other state-of-the-art approaches. Quantitative and qualitative evaluation show the superiority of 3D-RUNet and its ability to preserve the spectral fidelity of the enhanced HSI.**

***Index Terms*—Hyperspectral, Remote Sensing, Single Image Super Resolution, 3D Convolution, 3D-RUNet**

I. Introduction

Remote sensing technology has been rapidly developing since the 1980s, offering a wide range of applications related to vegetation, land cover land use, urbanization, and oceanography. The effectiveness of remote sensing imagery depends on its resolution, which can be categorized to spectral, spatial, and temporal. Only the first two categories are relevant to the scope of this paper. Spectral resolution measures the extent of the sensor's ability to capture wavelengths of the electromagnetic spectrum, while spatial resolution measures the smallest ground area that can be captured by a single pixel. An image that has high spectral resolution is referred to as Hyperspectral Image (HSI), while an image that has high spatial resolution is referred to as Multispectral Image (MSI). Due to sensor limitation, high spectral and spatial resolutions cannot be achieved simultaneously. This trade-off causes HSI to suffer from low spatial resolution. Overcoming this limitation is essential for the effective utilization of HSI in geological mapping, atmosphere monitoring, and mineral exploration, as enhancing the spatial resolution allows achieving image processing tasks, such as object detection and classification, with higher accuracy. Hence, researchers constantly strive to enhance the spatial resolution of HSI, a process called Super Resolution (SR). Generally, the approaches can be categorized into Fusion [1]–[3], and Single Image Super Resolution (SISR) [4]–[7]. Fusion approaches require auxiliary information, such as the sensor's Point Spread Function (PSF), which makes them impractical in scenarios where such information cannot be obtained. SISR approaches require no extra information other than the Low Resolution (LR) image that is being enhanced. Nonetheless, this convenience is consequently the reason why SISR is considered a notoriously ill-posed problem. Hence, SISR approaches are of particular interest in this study. Due to the success of Deep Convolutional Neural Networks (DCNNs) in image classification in 2014 [8], researchers studied DCNNs in the context of SISR, which prevailed over traditional methods, such as bicubic interpolation [9]. Currently, the literature is rich with DCNN-based SISR methods for both MSI and HSI [10]–[16]. Researchers argue that 2D DCNNs that are commonly used for MSI cannot be used for HSI enhancement due to the main difference between these two types of images; the spectral resolution [10]. 2D DCNNs fail to preserve the spectral fidelity of HSI. On the other hand, 3D DCNNs have proven effective in this area. For instance, Mei et al. [17] devised a 3D Full CNN (3D-FCNN) to mitigate the issue of spectral distortion when enhancing HSI. They test their algorithm on various standard datasets, including Pavia University, Washington DC, and Urban. The authors' experiments show that their network is superior against other 2D networks in terms of Peak Signal-to-Noise Ratio (PSNR), Structure Similarity Index Measurement (SSIM), and Spectral Angle Mapper (SAM).

One of the most famous 2D DCNNs that was developed for image segmentation tasks is UNet. This network can be re-purposed to achieve SISR. In fact, an improved version of UNet that achieves SISR was devised in [18], and it is referred to as Robust UNet (RUNet). It has been tested on RGB images and it proved superior to the original UNet, but there has been no attempt to utilize it for HSI-SR thus far. The main contribution of this paper is to develop, train, and study the effect of RUNet on HSI-SR. Additionally, RUNet will be extended to 3D to boost its performance and allow it to enhance HSI with minimal spectral distortions. The performance is compared against 3D-FCNN and other approaches using Salinas and Botswana datasets over scale factors 2 and 4. The comparison is performed quantitatively using PSNR, SSIM, and SAM, as well as qualitatively. The rest of the paper is organized as follows: Section II demonstrates the main approaches including the basic building blocks of encoders-decoders, Section III explores the datasets used, and lists and analyzes the results, and finally, Section IV summarizes and



(a) 2D

(b) 3D

Fig. 1: Visual illustration of 2D and 3D TC.

sets the future direction of this research.

II. Methodology

A. Encoder-Decoder Architecture

This type of architecture consists of two parts. The first one is the encoder, which extracts features from the input. The second part is the decoder, which translates the features taken from the encoder to estimate the original input. The encoder and decoder are mirror opposites of each other. This means that a convolutional layer in the encoder side is complemented with a Transpose Convolutional (TC) layer in the decoder side. Likewise, a pooling layer in the encoder side is complemented with an upsampling layer in the decoder side. This is done so that the output obtained would be of the same size as the input. The next subsections explain all the aforementioned operations.

1) 2D Convolution and 2D TC: A convolutional layer is a filter of several kernels that comprise input weights, and the output is a feature map of the input. A convolutional layer is typically followed by an activation function, such as Rectified Liner Unit (ReLU). For a single HSI band I of size $N \times N$ and a kernel K of size $M \times M$, 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 a subset of the HSI band that includes the pixel at location (x, y) 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. TC works the same way in principle. However, the kernel in equation 1 is replaced with the input band itself, which is convolved with a grid of the desired size, larger than the input, where the known input values are spread across the grid and the values in between are set to zero. While TC expands the spatial size of the input, convolution reduces it. To prevent spatial reduction, the input image must be padded at the borders. The simplest padding method is adding zeros to the boarders of the image [10].

2) 2D Pooling and 2D Upsampling: Pooling layer is another essential building block of CNNs that selects certain features and discards others depending on the type of pooling. The most commonly used one is max pooling, where a kernel passes through each band of the image to select the highest value within the kernel and discards the lower ones. On the other hand, upsampling layer replicates rows and columns values to achieve a desired spatial expansion in a manner similar to nearest neighbor interpolation.

B. 2D-UNet

UNet is a CNN that was originally built in 2015 for the purpose of segmenting biomedical images [19]. After its success, it was used for segmenting other types of images, including remote sensing imagery. UNet follows encoder-decoder architecture. UNet, and encoder-decoder networks generally, learn a one-to-one mapping between inputs and outputs. Therefore, a UNet that performs segmentation can be re-purposed to perform SISR. Instead of producing a binary mask with the same size as the input, it can produce an output of the same image type and size as the input. The work in [18] demonstrates UNet usage for MSI-SISR, particularly RGB images. Further, they introduce skip connections to the architecture to enhance its performance. Skip connections, also called residual connections, were introduced first in Residual Neural Network (ResNet) architecture [20]. They provide an alternative path from one layer to another one down the network, such that they add their respective outputs while skipping the layers in between them. The goal is to overcome vanishing gradient, an issue that becomes more prominent as the depth of the CNN increases. Introducing skip connections allows for better training of DCNNs without encountering vanishing gradients. The improved UNet architecture after adding skip connections is called RUNet, and it shows superiority over the original UNet and bicubic interpolation in terms of PSNR, SSIM, and Mean Squared Error (MSE). In this paper, RUNet is adapted for HSI-SR with minor modifications. For instance, all Batch Normalization (BN) layers are removed from the network, as it has been proven that this type of layer negatively affects SISR applications by adding artifacts [21]. Furthermore, the recreated version of the network is more compact since the training datasets are not large.

C. 3D-UNet

As discussed in Section I, 2D CNNs achieve decent performance in SISR of MSI, but they fall short when it comes to HSI due to their inability to preserve spectral signatures. A CNN has the potential to enhance HSI without spectral distortions of its building blocks can be extended to 3D. 3D convolution in principle is the same as 2D convolution, but instead of having a 2D filter, it has a 3D filter (cuboid) which includes the depth of the image in its calculations. For an HSI cube HS 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:

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

3D TC follows the same logic as the 2D one, but similar to 3D convolution, the filter is a cuboid. Figure 1 illustrates the differences between 2D and 3D TC. As for pooling, it splits the HSI into cuboid regions and computes the maximum value

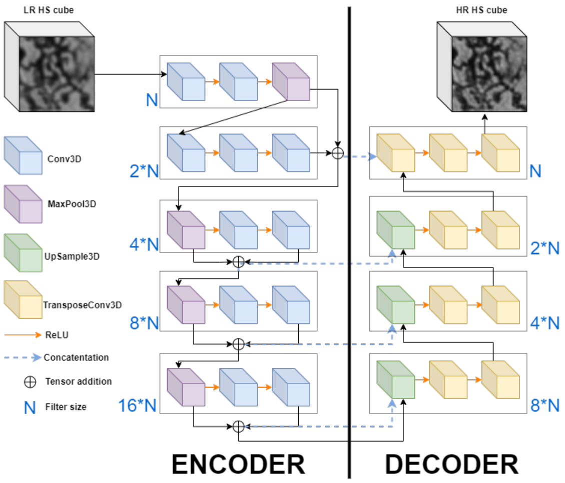


Fig. 2: Overall architecture of the proposed 3D-RUNet.

TABLE I: Description of the datasets used in this study.

Dataset	Sensor	Spatial Resolution	Bands	Size	Training	Testing
Salinas	AVIRIS	3.7m	204	512 × 217	21	3
Botswana	Hyperion	30m	145	1476 × 256	81	11

within each cuboid. Finally, 3D upsampling duplicates rows and columns values, in addition to depth values. The overall architecture of 3D-RUNet can be seen in Figure 2.

III. Results and Discussion

A. Datasets

The HSI datasets used in this study were captured using two different sensors. The first one is Airborne Visible/Infrared Imaging Spectrometer (AVIRIS), from which Salinas dataset was captured. The second one is Hyperion, from which Botswana dataset was captured. Training a DCNN using these datasets is a challenge because each dataset contains one image only, and training a DCNN requires a large amount of data. Therefore, each image is divided into patches of 64 × 64 pixels. Salinas and Botswana datasets provide an acceptable number of patches to train a DCNN. In order to generate an LR counterpart for each patch, it is downgraded using Gaussian blur, and then downsized by the required scale factor, which is upscaled again by the same scale factor using bicubic interpolation. The resulting degraded patch will be used as an input to the proposed 3D-RUNet as well as other DCNNs. Table I summarizes the spatial resolution, number of bands, the size, and the number of training and testing samples of each dataset.

B. Quantitative and Qualitative Evaluation

All networks have been developed, trained, and tested using Tensorflow-GPU library on NVIDIA Quadro P6000-24GB X 2 GPU with 380GB RAM. They were also trained under the same conditions. Each network has a learning rate of 10^{-5}, with Adam as an optimization function and 750 epochs. Early stopping strategy is followed, such that if the network does not improve after 10 consecutive epochs, the training stops. The training was repeated 5 times for each network and the average results were computed. Each network was tested on Salinas and Botswana datasets with scale factors 2 and 4. The results of Salinas dataset are summarized in Table II. The proposed 3D-RUNet prevails over the other methods in terms of PSNR, SSIM, and SAM for both scale factors, with the exception of SSIM for scale factor 4, where 3D-FCNN prevails by a margin of 0.01. As for Botswana dataset, the results show the superiority of 3D-RUNet against all other methods without exception, as seen in Table III. 3D-RUNet improvements are more evident when observing SAM, which shows its resilience in preserving spectral signature in both scale factors for both datasets. Additionally, qualitative results can be seen in Figure 3. Due to space limitation, only Salinas dataset results of scale factor 2 are displayed. It is evident that the 3D networks successfully captures the details of the image, while the 2D network introduces some artifacts in homogeneous areas. Furthermore, 3D-RUNet exhibits less blurriness than 3D-FCNN and appears closer to the original HR image. These results are consistent with the spectral signature of a target pixel plotted in Figure 4, where 2D-RUNET appears the least similar to the original HR signature, while 3D-RUNet follows the ground truth pattern more closely than the other methods.

TABLE II: Summary of Salinas dataset results.

Scale Factor	Method	PSNR (dB)	SSIM	SAM (°)
×2	Bicubic	36.26	0.95	2.87
	3D-FCNN	37.38	0.96	2.77
	2D-RUNet	38.69	0.96	2.16
	3D-RUNet	**39.15**	**0.97**	**2.15**
×4	Bicubic	31.66	0.90	5.25
	3D-FCNN	34.62	**0.93**	3.55
	2D-RUNet	34.21	0.91	3.57
	3D-RUNet	**34.65**	0.92	**3.47**

TABLE III: Summary of Botswana dataset results.

Scale Factor	Method	PSNR (dB)	SSIM	SAM (°)
×2	Bicubic	32.83	0.87	3.07
	3D-FCNN	35.12	0.92	2.39
	2D-RUNet	35.23	0.81	2.76
	3D-RUNet	**36.32**	**0.93**	**2.32**
×4	Bicubic	29.60	0.74	4.74
	3D-FCNN	30.16	0.79	4.59
	2D-RUNet	30.90	0.75	6.77
	3D-RUNet	**33.07**	**0.80**	**4.03**

IV. Conclusion

In this paper, 2D-RUNet was modified and extended to 3D. The network follows encoder-decoder architecture, which consists of layers that mirror each other to extract features and estimate the HR-HSI. 2D convolution and 2D max pooling in the encoder are complemented with 2D TC and 2D upsampling in the decoder, respectively. These operations are extended to 3D to obtain 3D-RUNet. The proposed network is compared to 2D-RUNet and 3D-FCNN over two different datasets, Salinas and Botswana, with scale factors 2 and 4. All networks are developed, trained, and tested under the same environment to ensure a fair comparison. Both qualitative and

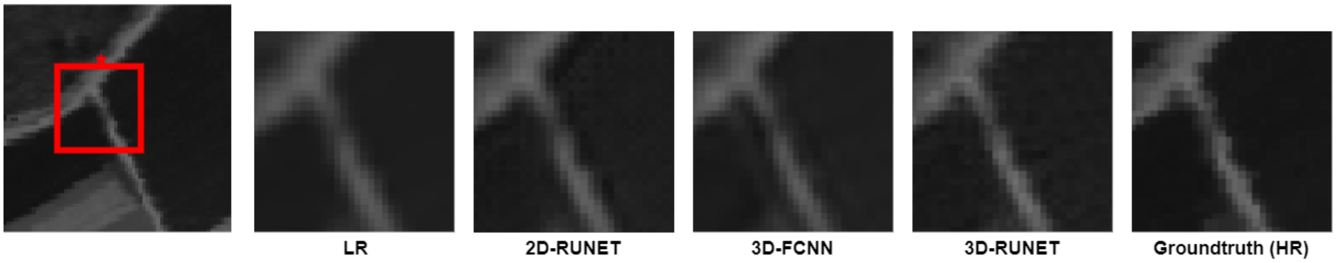


Fig. 3: Visual comparison between each output and the HR patch. The images are cropped from the 7th band of Salinas dataset for scale factor 2.

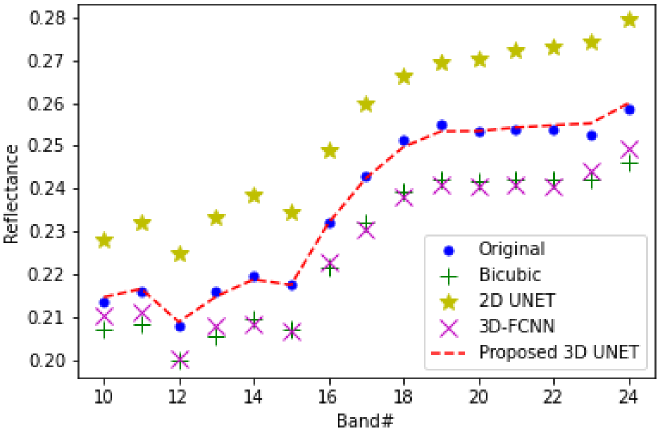


Fig. 4: Spectral signature comparison of each method at a target pixel. The plot is shown for bands 10-24 for clarity.

quantitative evaluations and comparisons in terms of PSNR, SSIM, and SAM indicate that 3D-RUNet outperforms other methodologies and minimizes spectral distortions.

References

[1] H. Irmak, G. B. Akar, and S. E. Yuksel, “Image fusion for hyperspectral image super-resolution,” in *2018 9th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 2018, pp. 1–5.

[2] G. Yu, F. Zhang, T. Hu, W. Li, and R. Tao, “Hyperspectral image super-resolution based on multiscale residual block and multilevel feature fusion,” in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 2170–2173.

[3] J. Chanussot, “On hyperspectral super-resolution,” in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 29–32.

[4] O. Sidorov and J. Y. Hardeberg, “Deep hyperspectral prior: Single-image denoising, inpainting, super-resolution,” in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 3844–3851.

[5] K. Zheng, L. Gao, B. Zhang, and X. Cui, “Multi-losses function based convolution neural network for single hyperspectral image super-resolution,” in *2018 Fifth International Workshop on Earth Observation and Remote Sensing Applications (EORSA)*, 2018, pp. 1–4.

[6] P. V. Arun, K. M. Buddhiraju, A. Porwal, and J. Chanussot, “Cnn-based super-resolution of hyperspectral images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 9, pp. 6106–6121, 2020.

[7] X. Wang, J. Ma, and J. Jiang, “Hyperspectral image super-resolution via recurrent feedback embedding and spatial–spectral consistency regularization,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, pp. 211–252, 2015.

[9] R. Keys, “Cubic convolution interpolation for digital image processing,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 6, pp. 1153–1160, 1981.

[10] N. Aburaed, M. Q. Alkhatib, S. Marshall, J. Zabalza, and H. Al Ahmad, “3d expansion of srcnn for spatial enhancement of hyperspectral remote sensing images,” in *2021 4th International Conference on Signal Processing and Information Security (ICSPIS)*, 2021, pp. 9–12.

[11] C. Dong, C. Change Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *CoRR*, vol. abs/1501.00092, 2015.

[12] J. Kim, J. K. Lee, and K. M. Lee, “Accurate image super-resolution using very deep convolutional networks,” *CoRR*, vol. abs/1511.04587, 2015.

[13] 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.

[14] 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.

[15] N. Aburaed, A. Panthakkan, S. Al Mansoori, 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.

[16] N. Aburaed, A. Panthakkan, M. Al-Saad, M. C. El Rai, S. Al Mansoori, H. Al-Ahmad, and S. 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.

[17] 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.

[18] X. Hu, M. A. Naiel, A. Wong, M. Lamm, and P. Fieguth, “Runet: A robust unet architecture for image super-resolution,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[19] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham, 2015, pp. 234–241, Springer International Publishing.

[20] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *CoRR*, vol. abs/1512.03385, 2015.

[21] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, “Esrgan: Enhanced super-resolution generative adversarial networks,” in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds., Cham, 2019, pp. 63–79, Springer International Publishing.