DeepSen3: Deep multi-scale learning model for spatial-spectral fusion of Sentinel-2 and Sentinel-3 remote sensing images

Ahed Alboody, Matthieu Puigt, Gilles Roussel, Vincent Vantrepotte, Cédric Jamet, Trung-Kien Tran

DEEPSEN3: DEEP MULTI-SCALE LEARNING MODEL FOR SPATIAL-SPECTRAL FUSION OF SENTINEL-2 AND SENTINEL-3 REMOTE SENSING IMAGES

Ahed Alboody[1], Matthieu Puigt[1], Gilles Roussel[1], Vincent Vantrepotte[2], Cédric Jamet[2], Trung-Kien Tran[2]

[1] Univ. Littoral Côte d'Opale, LISIC – UR 4491, F-62219 Longuenesse, France
[2] Univ. Littoral Côte d'Opale, CNRS, LOG – UMR 8187, F-62930 Wimereux, France

ABSTRACT

Recently, deep learning methods that integrate image features gradually became a hot development trend in fusion of multispectral and hyperspectral remote sensing images, aka multi-sharpening. Fusion of a low spatial resolution hyperspectral image (LR-HSI datacube) with its corresponding high spatial resolution multispectral image (HR-MSI datacube) to reconstruct a high spatial resolution hyperspectral image (HR-HSI) has been a significant subject in recent years. Nevertheless, it is still difficult to achieve a high quality of spatial and spectral information fusion. In this paper, we propose a Deep Multi-Scale Learning Model (called DeepSen3) of spatial-spectral information fusion based on multi-scale inception residual convolutional neural network (CNN) for more efficient hyperspectral and multispectral image fusion from ESA remote sensing satellite missions (Sentinel-2 and Sentinel-3 images). The proposed DeepSen3 fusion network was applied to Sentinel-2 MSI (13 spectral bands with a spatial resolution ranging from 10, 20 to 60 m) and Sentinel-3 OLCI (21 spectral bands with a spatial resolution of 300 m) images. Extensive experiments demonstrate that the proposed DeepSen3 network achieves the best performance (both qualitatively and quantitatively) compared with recent state-of-the-art deep learning approaches.

Index Terms— Deep Learning, Residual Convolutional Neural Network (ResNet-CNN), Multi-Scale Inception, Feature Extraction, Spatial-Spectral Image Fusion, Sentinel-2 and Sentinel-3 Remote Sensing Images, HyperSpectral Images (HSI), Multi-Spectral Images (MSI)

1. INTRODUCTION

Most remote sensing applications require images at the highest resolution both in spatial and spectral domains, which is impossible very hard to achieve by a single sensor. To alleviate this problem, many optical Earth observation satellites—such as Sentinel-2, Sentinel-3, Landsat 8, and MODIS—carry optical sensors, acquiring multi-modal data with different but complementary characteristics (spectral, spatial). In particular, a multi-spectral sensor acquires high spatial resolution images, while a hyperspectral sensor acquires low spatial resolution images with multiple bands (*datacube*). These modalities are known as hyper/multi-spectral images (*datacubes)*. In remote sensing applied to watercolor extraction and marine application, Sentinel-2 MSI and Sentinel-3 Ocean and Land Color Instrument (OLCI) are extensively used. Respectively, they have 13—with 10, 20, or 60 m spatial resolution—and 21 usable spectral bands—with 300 m spatial resolution—in the visible and near-infrared range. The technique to fuse these images is called multi-sharpening or image fusion [1, 2, 3]. It consists of combining relevant information from two or more MSI/HSI images into a single image with complementary spatial and spectral resolution characteristics. To that end, several methods have been proposed in the literature and are based on, e.g., component substitution—using, e.g., Gram-Schmidt [4], the Sylvester equation [5], or Principal Component Analysis [6]—coupled Nonnegative Matrix/Tensor Factorization (NMF/NTF)—e.g., [1, 2, 7, 8]—and more recently on deep learning, e.g., in [9]. Indeed, even if in [10], the authors suggested a method for the fusion of data from Sentinel-2 and Sentinel-3, they proposed in practice an approach to combine Sentinel-2 and Moderate Resolution Imaging Spectroradiometer (MODIS) images. The fusion of Sentinel-2 and Sentinel-3 data is let for future work in the conclusion of [10]. However, we have tested some of these methods to fuse Sentinel-2 MSI and Sentinel-3 OLCI images in [1].

With the rapid development of deep learning methods—especially convolutional networks (CNN) [11, 12, 13, 14]—these types of methods have become a growing trend in LR-HSI and HR-MSI image fusion [13, 16], super-resolution [15, 18, 19], and hyperspectral image pansharpening [14, 17, 20]. Deep learning methods show excellent performance on LR-HSI and HR-MSI fusion. It is still challenging to efficiently enhance the performance of image fusion in a learnable manner across new hyperspectral data sets such as Sentinel-2 and Sentinel-3, which is crucial for improving their fusion quality.

In order to provide more detailed spectral and spatial feature extraction for the spatial-spectral reconstruction hyperspectral images, we introduce multi-scale residual and inception convolutional blocks in feature extraction and one convolutional block in residual module which was used in super-resolution [15, 18, 19], pan-sharpening hyperspectral image [14, 17, 20] and image fusion [13, 16].

The remainder of the paper reads as follows. In Section 2, we briefly introduce our proposed method. Section 3 then presents experiments on real dataset and the reached multi-sharpening performance our method compared with recent state-of-the-art hyperspectral and

This work is partially funded by CNES within the TOSCA "OSYNICO" project. Experiments presented in this paper were carried out using the CALCULCO computing platform, supported by SCoSI/ULCO.

multispectral fusion image approaches. Finally, we conclude and discuss about future work in Section 4.

2. PROPOSED METHOD DEEPSEN3: DEEP MULTI - SCALE LEARNING MODEL FOR SPATIAL - SPECTRAL FUSION

In this paper, we establish a novel deep multi-scale residual and inception CNN network for hyperspectral and multispectral fusion image (see Figs. 1 and 2). The main contributions of our proposed method can be summarized as follows. Firstly, inspired by the ResNet [13, 15, 19, 20], multi-scale [14, 16] and Inception network architectures [14, 17, 18], we propose a deep multi-scale residual inception CNN network in our DeepSen3 method (see Figs. 2 and 3), with three modules of residual multi-scale and inception with feature extraction layers to reconstruct the desired HR-HSI (the fused image). Secondly, we introduce a new multi-scale residual inception CNN architecture to effectively boost the spatial-spectral feature extraction of the HR-HSI fused image. The multi-scale inception module consists of three branches with two convolution layers and the activation function *LeakyRelu.* Indeed, we use three local residual modules (refer to residual modules 1, 2, 3 in Fig. 3) for feature extraction and fusion feature maps and one residual module global to reconstruct the high quality of the HR-HSI fused image.

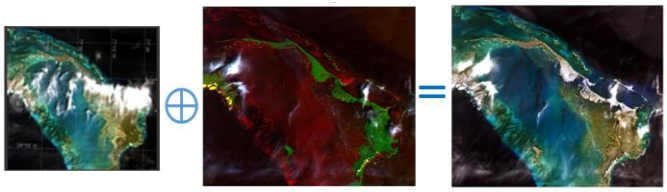


Fig. 1. Schematic diagram of hyperspectral and multispectral image fusion: hyperspectral resolution on an image from Sentinel-3 (LR-HSI), and multi-spectral image from Sentinel-2 (HR-MSI). The right image is the fused image HR-HSI.

2.1. Method Overview: DeepSen3

We design the DeepSen3 method to fuse two HR-MSI and LR-HSI datacubes with two distinct spatial and spectral resolutions.
Firstly, we up-sample hyperspectral images by the bi-cubic interpolation method. Then, we concatenate the up-sampled LR-HSI with HR-MSI to be the input of our proposed DeepSen3 network. DeepSen3 consists of three multi-scale inception and residual modules for feature extraction with only trainable convolution layers without pooling layers. These modules are divided into three main components: (i) multi-scale inception module for feature extraction and spatial-spectral fusion, (ii) residual modules for spatial-spectral fusion, and (iii) residual module global and multi-scale residual inception module for reconstruction image. In Fig. 3, we provide the overall structure of our proposed DeepSen3 method. The following section describes its network architecture in detail.

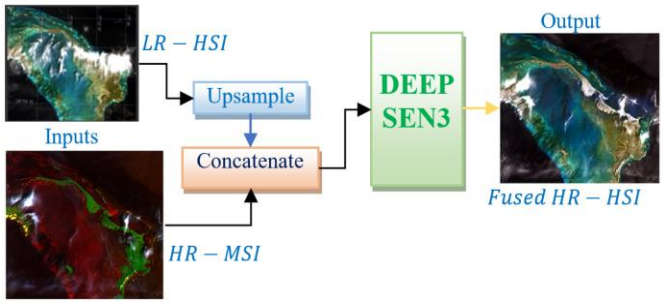


Fig.2. Workflow of the proposed method DeepSen3

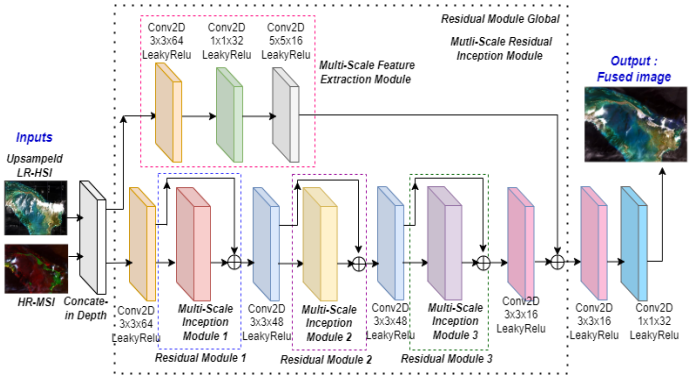


Fig. 3. DeepSen3 Architecture - Deep Multi-Scale Learning Model for Spatial-Spectral Fusion Remote Sensing Image

2.2. Multi-Scale Inception and Residual Modules for Feature Extraction

Following the principle of multi-scale residual block proposed in [13–20], and inspired by the hybrid dilated convolution (HDC) module proposed in [17] and Inception Block [18], we used three multi-scale Inception modules and three residual modules. Multi-Scale Inception module consists of three branches of two convolution layers with different kernel sizes is 3×3, 5×5 with different numbers (16, 32, 64 and 48) of channels to extract the feature maps of source images, where the activation function Leaky ReLU follows convolution layers. The feature maps of these convolution layers of multi-scale Inception modules are added through the residual module with identity mapping to generate the multi-scale residual inception features, and then to reconstruct the fused image.
In the multi-scale feature extraction stage, for inputs, a convolution layer with a kernel of size 3×3 is first used to extract low-level features of the source images. Convolution kernels of different sizes have different information perception capabilities and can extract detailed information at different scales. The convolution layers is followed by one common activation function Leaky ReLU (as shown in Fig. 4.). Leaky ReLU gives a better response than ReLU because it uses a learnable slope parameter instead of a constant slope parameter, which reduces the risk of over-fitting in the training. In order to obtain an accurate feature extraction map, it is necessary to take full advantage of the information of various scales in the input images. Therefore, in the part of the structure with residual

inception modules (as shown in Fig. 4.), we use three different branches to extract the feature information of different scales in the image. The sizes of the used convolution kernels are 3×3, and 5×5 with different channels (16, 32, and 64). Moreover, because the inception module can increase the width of the network and the adaptability to scales, we embed the multi-scale inception module in each branch of the residual module to improve the performance of the network fusion.

Figure 4 illustrates the multi-scale inception module's framework. On one hand, using the multi-scale inception module, convolution layers with different sizes of filters (such as 3×3 and 5×5) are used to extract detailed features at different scales. On the other hand, the 3×3, 5×5 convolutions in the multi-scale inception, and 1×1convolution and residual module global (as shown in Fig. 3) can superimpose more convolutions in the receptive field of the same size, and can extract richer features to fuse. At the same time, we use 1×1 convolution to reduce the dimensionality and the computational complexity. After extracting the feature information of different scales, we fuse them on the feature level and then use them as the input of the next convolutional layer. Figure 5 illustrates the main difference between residual blocks, inception module and multi-scale blocks with our Multi-scale-Residual-Inception Module proposed in our DeepSen3 architecture.

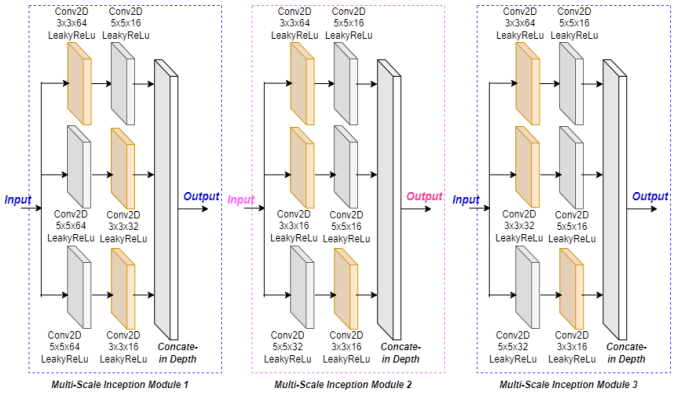


Fig.4. Multi-Scale Inception Modules for Spatial-Spectral Feature Extraction and Fusion

2.3 Model Implementation: Loss Function and Training Options

It is very important to choose a suitable loss function to accurately reconstruct the fused image. The mean square error loss (MSE) calculation is convenient and the convergence speed is fast. Therefore, we choose the mean square error loss as a loss function, which represents the average of the squares of deviations between the predicted fused image and reference (target). The MSE is calculated by the following formula:

$$Loss = \frac{1}{N}\sum_{i=1}^{N}\|Y_i - M_i\|^2,$$

where Y_i is the target HR-HSI reference, M_i is the predicted image by DeepSen3, N denotes the sample numbers in Y_i. We specify the size of the patch as 64 × 64 ×19 (width × height × (3 spectral bands of HR-MSI and 16 spectral bands of LR-HSI)).

In the training process, we train DeepSen3 using nine images of Sentinel-2 (HR-MSI with size of: Width=1792, Height=1792, and number of spectral band=3) with resolution 60 meters, and nine images Sentinel-3 (LR-HSI with size of: width=358, height=358, and 16 spectral bands) with resolution 300 meters for two coastal areas, i.e., Venice and Bahamas. We choose to randomly partition these images into 4032 image patches to train DeepSen3. The standard back propagation with *Adam* optimization algorithm (as optimizer) is utilized to minimize the loss function. *The initial learning rate of Adam* was set to $3*10^{-4}$, $\beta1=0.92$, $\beta2=0.95$, and the decay of the learning rate was set to 10^{-9}. *The mini-batch size is set to 16*.

The DeepSen3 model is implemented using *Python* programming language and *Keras deep learning library* with *TensorFlow* as computation backend. DeepSen3 model was trained *on Dell Precision Workstation 7540 with Intel CPU CORE i9 and GPU NVIDIA Quadro RTX3000*.

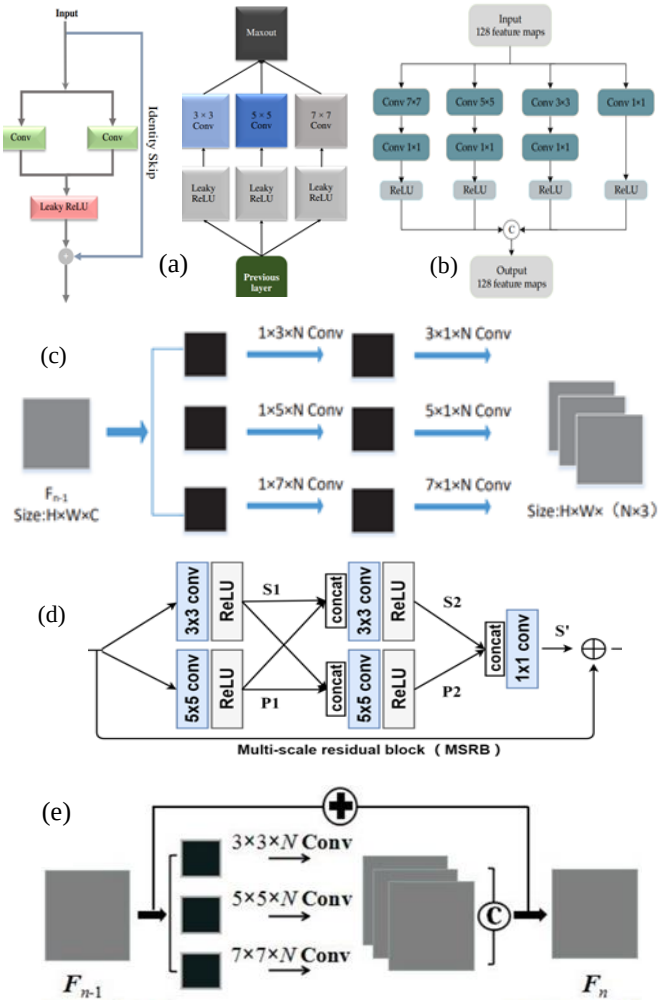
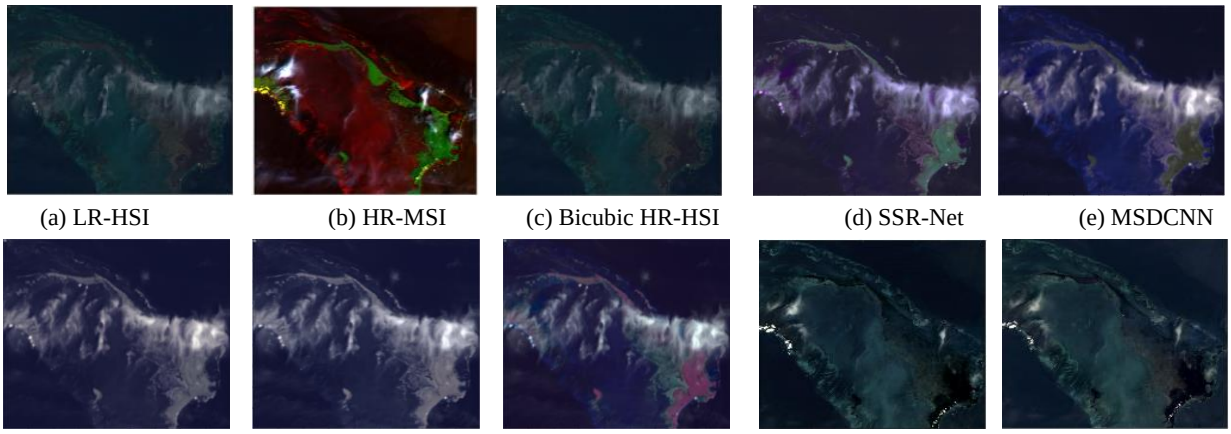


Fig. 5. Comparison: (a) Residual and Inception blocks [18], (b) Multi-scale feature extraction block [17], (c) Multi-Scale Asymmetric CNN [16], (d) Multi-scale residual block [15], (e) Multi-scale convolutional layer block with a short-distance skip connection [14], with our proposed modules (Multi-scale-Residual-Inception Module) in Fig. 4. and Fig .3.



(a) LR-HSI (b) HR-MSI (c) Bicubic HR-HSI (d) SSR-Net (e) MSDCNN

(f) TFNet (g) ResTFNet (h) ConSSFCNN (i) DeepSen3 (j) GT (HR-HSI)

Fig. 6: Fusion results of different methods on LR-HSI (Sentinel-3) and HR-MSI (Sentinel-2) dataset, where 'GT' represents the ground-truth image. Figure 6 shows the R-G-B images (8-6-4 bands) of the fused HR-HSI.

3. EXPERIMENTS & VALIDATION

We now investigate the performance reached by DeepSen3 and on state-of-the-art deep learning methods on real Sentinel-2 and Sentinel-3 datasets (Fig. 6). The selected study area is located in the islands of The Bahamas in the Atlantic Ocean. The Tongue of the Ocean is a deep-water basin in the Bahamas that is surrounded to the East, West, and South by a carbonate bank known as the Great Bahama Bank. The deep blue water of the Tongue is a stark contrast to the shallow turquoise waters of the surrounding Bank. Generally, waters that are optically shallow (e.g., Grand Bahama Bank) appear blue–green due to high bottom reflectance contributions while optically deep waters appear dark blue. The center of the Bahamas image is located at the following coordinates: Lat-26◦ 35'58.50"N, and Lon-77◦ 28'29.93"W (DMS), Projection UTM, Zone 18 N, and World Geodetic System 1984.

In this study, we use one Sentinel-2 (S2B-MSI) and one Sentinel-3 (S3A-OLCI) images taken on the same date (April, 22nd 2020) to evaluate DeepSen3. However, Sentinel-2 already has 3 different spatial resolution at different wavelengths. One must thus choose a target spatial resolution among the three available, i.e., 10, 20, or 60 m. For these data, for watercolor extraction and marine application, we choose to fuse 16 spectral bands from Sentinel-3 (spatial resolution 300 meters) and 3 spectral bands of Sentinel-2 (resolution 60 meters). In addition, we have the size of the patch is 64×64×19 (width × height × (3 spectral bands of HR-MSI/Sentinel-2 and 16 spectral bands of LR-HSI /Sentinel-3)).

In this section, we evaluate the performance of hyperspectral and multispectral fusion methods. To that end, we consider classical and modern multi-sharpening techniques based on deep learning approaches in [11–14, 19, 20] for multi-/pan-sharpening, fusion and super-resolution hyperspectral image as in [1, 2, 3]. To that end, traditional methods of CNMF and others tested in [1, 2, 7], and deep learning methods such as SSR-NET [11], SSFCNN and ConSSFCNN [12], TFNet and ResTFNet [13], and MSDCNN [14] are selected as the comparison approaches to evaluate the performance of our proposed method DeepSen3 (see Table 1, and Figure 6). For the traditional methods, except data processing, all the parameters are set as the same as in the original literature. For all the deep learning models, the number of input and output channels are adapted to Sentinel-2 and Sentinel-3 datasets.

In order to assess the sharpening performance, we use some classical quantitative measures [1, 2, 11, 12], i.e., (i) the Peak Signal-to-Noise Ratio (PSNR) in dB— which is the ratio between the highest possible signal energy and the noise energy—(ii) the Spectral Angle Mapper (SAM) in radian—which is a pixel wise measure of the angle (converted from degrees to radians) between the reference spectrum and the fused one. SAM values near zero indicate local high spectral quality and we use the average SAM value with respect to pixels for the quality index of the entire data set—(iii) the ERGAS measure, i.e., a normalized average error of each band of the processed image.

According to the considered experiments presented in Figure 6 and Table 1, DeepSen3 has a better performance compared with five deep learning methods (SSRNet, ConSSFCNN, MSDCNN, TFNet and ResTFNet. One advantage of our DeepSen3 method is that it can partially remove clouds from Sentinel-3, as also found in [1] for some non-deep-learning-based methods, i.e., CNMF and GSA.

Perf. obtained with 60 m spatial resolution			
Method	**PSNR (dB)**	**SAM (radian)**	**ERGAS**
ConSSFCNN	28.9	3.84	7.90
ResTFNet	27.8	4.58	13.35
TFNet	27.9	4.67	9.37
MSDCNN	**32.6**	**2.67**	**7.41**
SSR-NET	**32.5**	**2.42**	**7.40**
DeepSen3	**43.8**	**0.11**	**5. 56**

TABLE 1. Comparison metric results of the proposed method DeepSen3 and state-of-the-art methods (SSR-NET [11], ConSSFCNN [12], TFNet and ResTFNet [13], and MSDCNN [14]) on Sentinel-2 and Sentinel-3 dataset.

4. CONCLUSION AND PERSPECTIVES

In this paper, we presented a deep multi-scale learning model to fuse Sentinel-2 and Sentinel-3 images. Based on the existing multi-/pan-sharpening and super-resolution hyperspectral deep learning approaches, we designed multi-scale residual inception blocks to fuse these images in order to extract richer and more complete spatial and spectral information. A high quality fused image can be obtained with full consideration of different spectral and spatial characteristics. According to the considered experiments, DeepSen3 provides a better performance compared with state-of-art deep learning methods.

In future work, we hope to propose a Deep Convolutional Generative Adversarial Networks (GAN) to fuse Sentinel-2 and Sentinel-3 images, and to compare it with an extension of DeepSen3 method to be able to take into account 11 bands of Sentinel-2 with atmospheric correction in order to provide a new super-resolution Sentinel-3 image which will also be atmospherically corrected.

5. REFERENCES

[1] A. Alboody, M. Puigt, G. Roussel, V. Vantrepotte, C. Jamet and T. K. Tran, "Experimental Comparison of Multi-Sharpening Methods Applied To Sentinel-2 MSI and Sentinel-3 OLCI Images," Proc. IEEE WHISPERS'21, pp. 1-5.

[2] N. Yokoya, C. Grohnfeldt, and J. Chanussot, "Hyperspectral and multispectral data fusion: A comparative review of the recent literature," IEEE Geosci. Remote Sens. Mag., vol. 5, no. 2, pp. 29–56, 2017.

[3] L. Loncan, L. B. De Almeida, J. M. Bioucas-Dias, X. Briottet, J. Chanussot, N. Dobigeon, S. Fabre, W. Liao, G. A. Licciardi, M. Simoes, et al., "Hyperspectral pansharpening: A review," IEEE Geosci. Remote Sens. Mag., vol. 3, no. 3, pp. 27–46, 2015.

[4] C. A. Laben and B. V. Brower, "Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening," 2000, US Patent 6,011,875.

[5] Q. Wei, N. Dobigeon, and J.-Y. Tourneret, "Fast fusion of multi-band images based on solving a sylvester equation," IEEE Trans. Image Process., vol. 24, no. 11, pp. 4109–4121, 2015.

[6] P. Kwarteng and A. Chavez, "Extracting spectral contrast in landsat thematic mapper image data using selective principal component analysis," Photogramm. Eng. Remote Sens, vol. 55, no. 1, pp. 339–348, 1989.

[7] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," IEEE Trans. Geosci. Remote Sens., vol. 50, no. 2, pp. 528–537, 2011.

[8] S. Li, R. Dian, L. Fang, and J. M. Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," IEEE Trans. Image Process., vol. 27, no. 8, pp. 4118–4130, 2018.

[9] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 2, pp. 295–307, 2015.

[10] A. Korosov and D. Pozdnyakov, "Fusion of data from Sentinel-2/MSI and Sentinel-3/OLCI," in Living Planet Symposium, 2016, vol. 740, p. 405.

[11] X. Zhang, W. Huang, Q. Wang and X. Li, "SSR-NET: Spatial–Spectral Reconstruction Network for Hyperspectral and Multispectral Image Fusion," IEEE Trans. Geosci. Remote Sens., vol. 59, no. 7, pp. 5953-5965, July 2021.

[12] X.-H. Han, B. Shi, and Y. Zheng, "SSF-CNN: Spatial and spectral fusion with cnn for hyperspectral image super-resolution," in Proc. IEEE ICIP'18, 2018, pp. 2506–2510.

[13] X. Liu, Q. Liu, and Y. Wang, "Remote sensing image fusion based on two-stream fusion network," Information Fusion, vol. 55, pp. 1–15, 2020.

[14] Q. Yuan, Y. Wei, X. Meng, H. Shen and L. Zhang, "A Multi-scale and Multi-depth Convolutional Neural Network for Remote Sensing Imagery Pan-Sharpening," in IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 11, no. 3, pp. 978-989, March 2018.

[15] L. Juncheng, F. Fang, K. Mei and G. Zhang, "Multi-scale Residual Network for Image Super-Resolution," in Proc. ECCV, 2018.

[16] J. Wei, Y. Xu, W. Cai, Z. Wu, J. Chanussot and Z. Wei, "A Two-Stream Multiscale Deep Learning Architecture for Pan-Sharpening," in IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 13, pp. 5455-5465, 2020.

[17] L. Weisheng, X. Liang, and M. Dong, "MDECNN: A Multiscale Perception Dense Encoding Convolutional Neural Network for Multispectral Pan-Sharpening," Remote Sensing 13, no. 3: 535, 2021.

[18] Muhammad, W., Bhutto, Z., Ansari, A., Memon, M.L., Kumar, R., Hussain, A., Shah, S.A.R., Thaheem, I., Ali, S, "Multi-Path Deep CNN with Residual Inception Network for Single Image Super-Resolution," Electronics, 2021.

[19] J. -F. Hu, T. -Z. Huang, L. -J. Deng, T. -X. Jiang, G. Vivone and J. Chanussot, "Hyperspectral Image Super-Resolution via Deep Spatiospectral Attention Convolutional Neural Networks," in IEEE Trans. Neural Netw. Learn. Syst, pp.1-15, 2021.

[20] L. He, J. Zhu, J. Li, A. Plaza, J. Chanussot and B. Li, "HyperPNN: Hyperspectral Pansharpening via Spectrally Predictive Convolutional Neural Networks," in IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 12, no. 8, pp. 3092-3100, Aug. 2019.