A PIXEL LEVEL SCALED FUSION MODEL TO PROVIDE HIGH SPATIAL-SPECTRAL RESOLUTION FOR SATELLITE IMAGES USING LSTM NETWORKS

Carlos A. Theran[1,2], *Michael A. Álvarez*[1,3], *Emmanuel Arzuaga*[1,2,3], *Heidy Sierra*[1,2]

1. Laboratory for Applied Remote Sensing, Imaging and Photonics
2. Department of Computer science & Engineering
3. Department of Electrical and Computer Engineering
University of Puerto Rico Mayaguez

ABSTRACT

Pixel-level fusion of satellite images coming from multiple sensors allows for an improvement in the quality of the acquired data both spatially and spectrally. In particular, multispectral and hyperspectral images have been fused to generate images with a high spatial and spectral resolution. In literature, there are several approaches for this task, nonetheless, those techniques still present a loss of relevant spatial information during the fusion process. This work presents a multi scale deep learning model to fuse multispectral and hyperspectral data, each with high-spatial-and-low-spectral resolution (HSaLS) and low-spatial-and-high-spectral resolution (LSaHS) respectively. As a result of the fusion scheme, a high-spatial-and-spectral resolution image (HSaHS) can be obtained. In order of accomplishing this result, we have developed a new scalable high spatial resolution process in which the model learns how to transition from low spatial resolution to an intermediate spatial resolution level and finally to the high spatial-spectral resolution image. This step-by-step process reduces significantly the loss of spatial information. The results of our approach show better performance in terms of both the structural similarity index and the signal to noise ratio.

Index Terms— Data Fusion, Long Short Term Memory, Pixel level, Super resolution, hyperspectral image, multispectral image.

1. INTRODUCTION

The availability of data captured by different remote sensing instruments has been increasing, opening possibilities to create new processing approaches for classification, enhancement, and tracking of features of any captured signal. Currently, the fusion of satellite images coming from multiple sensors has gain relevant attention, particularly in the development of new techniques aiming to improve spatial features by generating high-resolution images [1]. Image fusion is typically divided into three different levels of details; pixel level fusion, feature level fusion and decision level fusion [2], particularly pixel level fusion has gained substantial interest for multispectral (MS) and hyperspectral (HS) images. In recent studies, sparse representation methods have been proposed in order to generate optimum spatial-spectral resolution. Such a problem has been formulated as an inverse problem whose solution is the target image, represented by atoms of dictionaries [3] or using sparse matrix factorization schemes [4].

Machine learning algorithms have been successfully used for the analysis of remote sensing data in different applications. For example, neural network models have been proposed for super-resolution [5] and support vector machines in Pan-sharpening methods [6]. Moreover, deep learning has been adopted for super-resolution [7, 8] with excellent results. These approaches create an end-to-end mapping between low-resolution and high-resolution images. However, this mapping generates a loss of information (spatial features) when they generate high-resolution images. In this work, we propose a scalable-fusion model in order to reduce the loss of information, this approach learns how to transition from low resolution to an intermediate resolution stage and finally to a high resolution result.

Our computational approach called scaled-fusion, fuses two types of data; 1. Multispectral image with high-spatial resolution and low-spectral resolution (HSaLS), and 2. Hyperspectral images with low-spatial resolution and high-spectral resolution (LSaHS). As a result, the model provides a HSaHS image. Moreover, this model will cover the loss of information during the process of getting an HSaHS image by scalable learning. This means that if an image has 8m of spatial resolution and the enhancement target is 1m data, our model learns how to go from 8m to 4m and finally from 4m to 1m. This model is based on the best characteristic of long short term memory (LSTM) networks, which is to learn from past observations.

The separability of spectral and spatial information into the fusion of multispectral and hyperspectral images allows for

This material is based upon work supported by the National Science Foundation under Grant No. OAC-1750970.

the analysis for feature space and reduction along the spectral dimension. Different methods have been proposed, such as component substitution, which consists of transforming the spectral information into another feature space. Consequently, separate spatial and spectral informations can be obtained. Typical principal components analysis, singular value decomposition (SVD), and GramSchmidt orthonormalization [9] are well known for separability of spectral and spatial information. For the purpose of minimizing the loss of information, as a result from the dimension reduction process, our approach performs an SVD transformation [9]. The rest of the paper is organized as follows. Section 2 discuses the fusion problem and the new approach within the proposed framework. Section 3 discusses simulation results using different metrics, and the conclusions are reported in section 4.

2. THEORETICAL FRAMEWORK OF SVD & LSTM

The efficiency of our approach lies in the implementation of the LSTM model for spatial enhancement preserving spatial content, the following section briefly describes the SVD and LSTM methods and how they are employed in our proposed model.

2.1. Singular Value Decomposition

Hyperspectral sensors capture hundreds of spectral bands, this fact hinders the removal of redundant information that does not represent an improve to the analysis to scene of study. In our approach we need to reduce the spectral dimensionality of the hyperspectral images in order to remove redundant spectral bands and keep only the bands that represent the majority of information of the scene. For this purpose, we use the SVD as a technique that allows us separate the spatial and spectral information. Moreover, the SVD let us discriminate those bands with low information. Given a hyperspectral image $A \in \mathbf{R}^{p\times b}$ where b is the number of bands and p is the number of pixels. The SVD applied to A, gives the following factorization:

$$A = USV^T$$

where the matrix $U \in \mathbf{R}^{p\times p}$ is an orthogonal matrix, whose columns are eigenvectors of pixels, $S \in \mathbf{R}^{p\times b}$ is a diagonal matrix of which elements are the singular values that represent the energy of each pixel by bands. In this decomposition the matrix $\Gamma = US$ represents the spatial information and V contain the spectral information contents.

2.2. Long Short Term Memory Network

Our proposed fusion scheme relies on LSTM networks [10]. The idea behind LSTM networks is the inclusion of a self-loop that helps the gradient flow for each layer, resulting in the capability of remembering information for long periods of time. The weight of each self-loop is controlled by a hidden layer. This method requires more parameters and a system that controls the flow of information (cell stage), regulated by structures called gates. Lets us define each of the gates that compound a LSTM. The **forget gate** defines which characteristic of the past information we will keep for the actual prediction, this gate is defined as follows for a time $t \in \mathbf{R}$:

$$f_i^{(t)} = \sigma\left(b_i^f + \sum_j U_{i,j}^f x_j^{(t)} + \sum_j W_{i,j}^f h_j^{t-1}\right) \quad (1)$$

where $x^{(t)}$ is the actual input vector, the output of the LSTM at the current hidden layer is $\mathbf{h}^{(t)}$, and $\mathbf{b}^{(f)}$, $\mathbf{U}^{(f)}$, $\mathbf{W}^{(t)}$ are biases, input weights, and recurrent weights respectively. The next state decides the new information that will be stored in the cell state c_t. For this purpose, we need to find values that update $i_i^{(t)}$ as well as create a vector of new candidate values $\hat{c}_t$. The computation of $i_i^{(t)}$ and $\hat{c}_t$ is given by the following equations:

$$i_i^{(t)} = \sigma\left(b_i + \sum_j U_{i,j} x_j^{(t)} + \sum_j W_{i,j} h_j^{t-1}\right) \quad (2)$$

and

$$\hat{c}_i^{(t)} = \tanh\left(b_i^c + \sum_j U_{i,j}^c x_j^{(t)} + \sum_j W_{i,j}^c h_j^{t-1}\right) \quad (3)$$

where $\mathbf{b}^{(c)}$, $\mathbf{b}$, $\mathbf{U}^{(c)}$, $\mathbf{U}$, $\mathbf{W}^{(c)}$, $\mathbf{W}$ are biases, input weights, and recurrent weight. The new cell state c_t is calculated using the information computed at this point.

$$c_i^{(t)} = f_i^{(t)} \times c_i^{(t-1)} + i_i^{(t)} \times \hat{c}_i^{(t)} \quad (4)$$

In equation (4) there is a piece-wise operation. The output of the actual hidden layer $h^{(t)}$ is based on the cell state $c_i^{(t)}$, but it can be filtered or turned off by the output gate o_i^t. Both formulas are described bellow.

$$o_i^t = \sigma\left(b_i^o + \sum_j U_{i,j}^o x_j^{(t)} + \sum_j W_{i,j}^o h_j^{t-1}\right) \quad (5)$$

$$h^{(t)} = \tanh(c_i^t) o_i^t \quad (6)$$

where $\mathbf{b}^{(o)}$, $\mathbf{U}^{(o)}$, $\mathbf{W}^{(o)}$ are biases, input weights, and recurrent weight. LSTM have proven to learn long-term dependencies. With this architecture we will retain information from low resolution to intermediate resolution in order to get a high resolution. Figure 1 presents the proposed model for high spatial resolution images based in LSTM.

The complete training procedure is illustrated in Algorithm 1. The LSTM has the capability of remembering past events. Using this advantage we can learn how to go from low spatial resolution to high spatial resolution using an intermediate resolution. Consequently, the information loss between low resolution and high resolution can be minimized.

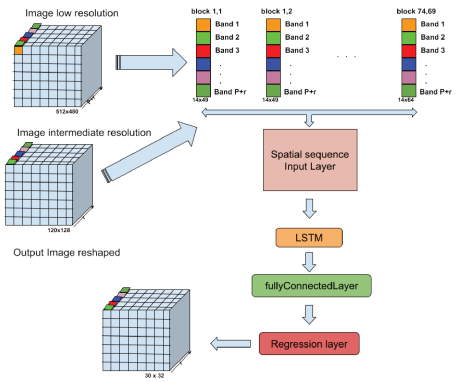

Fig. 1. LSTM architecture for high spatial resolution images

Algorithm 1: SVD & LSTM training program

Input : A trainig matrices: $M_H \in \mathcal{R}^{p_1 \times b}$: HSaLS, $H_1 \in \mathcal{R}^{p_2 \times q}$, $H_2 \in \mathcal{R}^{p_3 \times q}$: LSaHS
s.t. $p_3 < p_2 < p_1$, $b < q$

Compute for each $H_{i \in [1,2]}$: $[U_i, S_i, V_i] = \text{svd}(H_i)$;
Define: $\Gamma_i = U_i \cdot S_i$
for $r \in$ [6 10 15 20] & $i =$ [1 2] **do**
  $\Gamma_i r = \Gamma_i(:, 1:r)$
  Reshape $\Gamma_1 r$, $\Gamma_2 r$ to $3D$ format, and apply low pass filter:
  $\Gamma_i rd = \text{lowFilter}\{\Gamma_i r\}$
  Decimation in Multi-Spectral data M_H:
  $M_{Hi} = \text{decimationFilter}\{M_H$, $\text{size}\{\Gamma_i rd\}\}$
  Create input for training LSTM network:
  $X_{train}\{i\} = \text{cat}\{M_{Hi}$, $\Gamma_i rd\}$
  Create target for training LSTM network:
  $\Gamma_{train}\{i\} = \Gamma_i r$
  $net\{r\} = \text{LSTM\_Train}\{X_{train}$, $\Gamma_{train}\}$
end
Output: A set of neural networks trained: net

3. EXPERIMENTAL EVALUATION

3.1. Datasets

The sets of data used in this work consist of 3 different hyperspectral images: Salinas, Indian Pines and Enrique Reef.The Salinas hyperspectral image has a resolution of 3.7m, 204 bands and was collected by the AVIRIS sensor. It consists of 512×217 pixels. There are 16 classes: Brocoli green weeds 1, Brocoli green weeds 2, Fallow, Fallow rough plow, Fallow smooth, Stubble, Celery, Grapes untrained, Soil vineyard develop, Corn senesced green weeds, Lettuce romaine 4wk, Lettuce romaine 5wk, Lettuce romaine 6wk, Lettuce romaine 7wk, Vineyard untrained, and Vineyard vertical trellis.

The Indian Pines hyperspectral image was gathered by the AVIRIS sensor, consisting of 145×145 pixels and 224 spectral bands in the wavelength range 400 to 2500 nm. The number of bands were reduced to 200 by removing high water absorption bands. This scene has 16 classes: Alfalfa, Corn-notil, Corn-mintil, Corn, Grass-pasture, Grass-trees, Grass-pasture-mowed, Hay-windrowed, Oats, Soybean-notill, Soybean-mintill, Soybean-clean, Wheat, Woods, Buildings-Grass-Trees-Drives, and Stone-Steel-Towers.

There are two image groups from the Enrique Reef dataset. The first group consists of one high spatial resolution image taken from the multispectral sensor of IKONOS, such image was acquired in 2005. The image contains four layers: red, blue, green and near infrared, the spatial resolution of this data is 1m. There are 6 classes: Mangrove, Deep water, Coral, Sand, Sea grass, and Flat reef. The second group of images consist of images taken from the AISA Eagle sensor with a spatial resolution of 1, 2, 4, and 8m. The images were captured by the Galileo group in 2007. The number of bands for the images taken by the AISA Eagle sensor is 128. For our tests we used the high resolution hyperspectral image (1m), and the low resolution hyperspectral image (8m).

3.2. Metrics

In order to evaluate the performance of our method for fusion, two metrics were selected. The structural similarity index (SSIM)[11] given in the equation (7), power signal noise ratio (PSNR)[12] given in the equation (8).

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(\sigma_{xy} + c2)}{\mu_x^2 + \mu_y^2 + c2} \tag{7}$$

where μ_x and μ_y are the average of x and y respectively. σ_x^2 and σ_y^2 are the variance of x and y respectively, and $c_1 = (k_1, L)^2$, $c_2 = (k_2, L)^2$ where $k_1 = 0.01, k_2 = 0.03$ and L the dynamic range of the pixel-values. Now the PSNR modeled by

$$PSNR = 10 \log \left(\frac{R^2}{RMSE} \right) \tag{8}$$

where RMSE is the well known Root mean square error formula and here R is the maximum fluctuation in the input image. A high SSIM when compared to a high resolution reference image in our case means that a better HSaHS image was obtained. By other hand, PSNR compares the level of a desired signal to the level of background noise. Thus a higher PSNR a means that there is more useful content in the obtained data.

3.3. Results and Discussion

In this section we present the results obtained applying the proposed model. Using the metrics mentioned in section 3.2, we show in the tables 1, 2 and 3 the performance of two different deep leaning models: the first one is our LSTM model described in section 2.2 and the second is CNN proposed by Palsson [8]. For CNN we create two cases: CNN 4-1 and CNN 8-1. CNN 4-1 is a trained model to generate a HSaHS image from the input-pair (HSaLS LSaHS) where LSaHS is scaled 1/4. Similar to CNN 4-1, CNN 8-1 is a trained model, but the LSaHS image is scaled 1/8.

The datasets of Salinas and Indian Pines are augmented by the construction of one HSaLS and two LSaHS images; the HSaLS image is obtained by average of the bands in the wavelengt range blue (445 to 516 nm), green (506 to 595 nm), red (632 to 698 nm) and near IR (757 to 853 nm), the LSaHS images are generated using bicubic decimation process for two scaled factors, 1/4 and 1/8, w.r.t. the inititial resolution that is 3.7 meter.

A sample of the resultant dataset is shown in Fig. 2(a,b,c,d) and Fig. 3(a,b,c,d), for Salinas and Indian Pines respectively. To reach 80% of the information with randomly chosen samples we select 3800, 750, and 8200 patches for Salinas, Indian Pines and Enrique, respectively. To control the sequence of patches and make repeatable the tests a normally distributed random number generator is used. The test is performed 4 times in each model, and the average value is shown in each case:

- Table 1 shows the performance of the methods using the augmented Salinas dataset. A sample of the reconstruction of LSTM and CNN 4-1 is shown in Fig. 2(e) and (f), respectively.
- Table 2 shows the performance of the methods using the augmented Indian dataset. A sample of the reconstruction of LSTM and CNN 4-1 is shown in Fig. 3(e) and (f), respectively.
- Table 3 shows for the performance using the Enrique dataset. A sample of the reconstruction of LSTM and CNN 4-1 is shown in Fig. 4(e) and (f), respectively.
- Table 4 shows for the performance using the HSaLS image of Enrique dataset and generated LSaHS image from the HSaHS image of Enrique dataset.

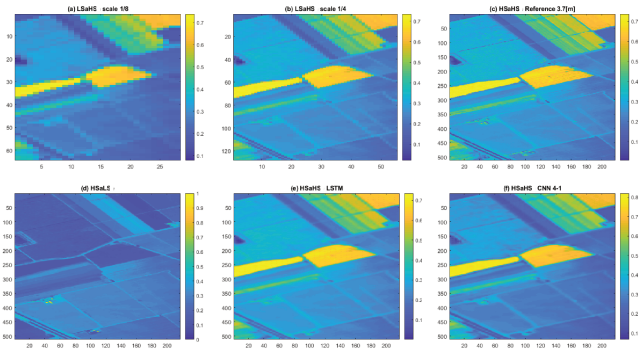

Fig. 2. Experiment with Salinas dataset, (a) LSaHS image of scale 1/8. (b) LSaHS image of scale 1/4. (c) Reference LSaHS data, resolution 3.7[m]. (d) HSaLS image. (e) reconstruction HSaHS performed with the LSTM model. (f) reconstruction HSaHS performed with the CNN model.

As can be observed from the tables, the LSTM approach showed the higher SSIM, PSNR and lower RMSE values for all experiments in all image datasets. These results provide confidence that the approach is capable of preserving better spatial features, producing a higher spatial resolution image.

# bands		6	10	15	20
SSIM	LSTM	0.958	0.958	0.959	0.959
	CNN 4-1	0.957	0.957	0.957	0.957
	CNN 8-1	0.9288	0.9301	0.9309	0.9321
RMSE	LSTM	0.012	0.012	0.012	0.012
	CNN 4-1	0.017	0.017	0.017	0.017
	CNN 8-1	0.0202	0.0199	0.0197	0.0196
PSNR	LSTM	35.98	36.04	36.06	36.08
	CNN 4-1	33.34	33.34	33.37	33.38
	CNN 8-1	31.80	31.93	32.01	32.04

Table 1. Performance of LSTM and CNN models w.r.t. bands number for Salinas dataset.

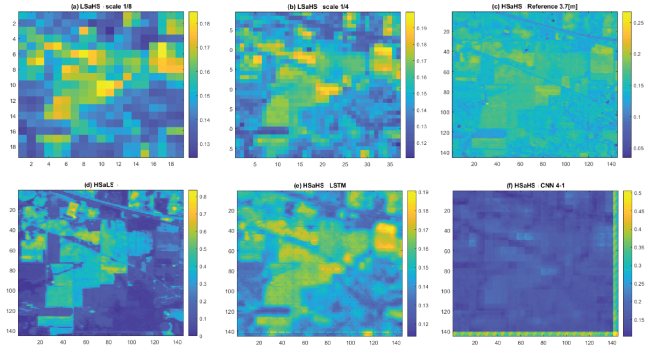

Fig. 3. Experiment with Indian Pines dataset, (a) LSaHS image of scale 1/8. (b) LSaHS image of scale 1/4. (c) Reference LSaHS resolution 3.7[m]. (d) HSaLS image. (e) reconstruction HSaHS performed with the LSTM model. (f) reconstruction HSaHS performed with the CNN model

# bands		6	10	15	20
SSIM	LSTM	0.910	0.903	0.901	0.901
	CNN 4-1	0.823	0.815	0.810	0.812
	CNN 8-1	0.691	0.693	0.687	0.694
RMSE	LSTM	0.020	0.020	0.020	0.020
	CNN 4-1	0.079	0.087	0.093	0.091
	CNN 8-1	0.166	0.159	0.162	0.153
PSNR	LSTM	31.81	31.76	31.69	31.73
	CNN 4-1	19.44	18.55	18.00	18.19
	CNN 8-1	13.03	13.44	13.23	13.76

Table 2. Performance of LSTM and CNN models w.r.t. bands number for Indian Pines dataset.

# bands		6	10	15	20
SSIM	LSTM	0.817	0.817	0.817	0.817
	CNN 4-1	0.730	0.730	0.729	0.729
RMSE	LSTM	0.024	0.023	0.023	0.023
	CNN 4-1	0.059	0.059	0.059	0.059
PSNR	LSTM	32.51	32.57	32.56	32.55
	CNN 4-1	24.25	24.24	24.24	24.23

Table 3. Performance of LSTM and CNN models w.r.t. bands number for Enrique Reef dataset.

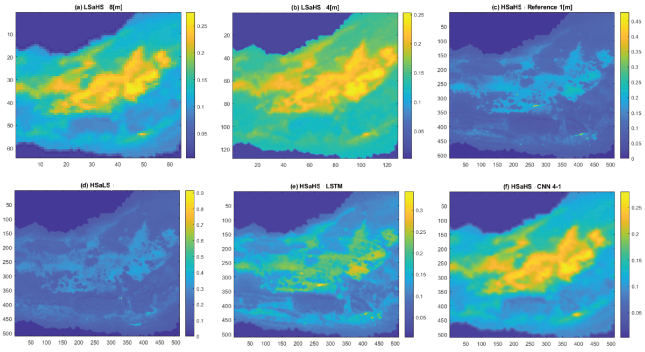

Fig. 4. Experiment with Enrique Reef dataset, (a) HS image of resolution 8m. (b) HS image of resolution 4m. (c) Reference LSaHS resolution 1m. (d) HSaLS image of 1m. (e) reconstruction HSaHS performed with the LSTM model. (f) reconstruction HSaHS performed with the CNN model

# bands		6	10	15	20
SSIM	LSTM	0.972	0.974	0.973	0.973
	CNN 4-1	0.956	0.955	0.955	0.955
RMSE	LSTM	0.009	0.009	0.009	0.009
	CNN 4-1	0.031	0.031	0.031	0.031
PSNR	LSTM	40.56	40.81	40.77	40.77
	CNN 4-1	29.97	29.92	29.90	29.89

Table 4. Performance of LSTM and CNN models w.r.t. bands number for simulated Enrique Reef dataset.

4. CONCLUSION

The proposed model has shown better results using a scalable learning approach for four different data set and metrics, as can be seen in the tables presented in section 3.3. Also, we have proof numerically that the LSTM is a useful architecture to generate HSaHS images compared with other architecture in literature. The numerical result with a few bands has achieved good reconstruction. Particularly, the higher SSIM were reached in band 6, 20 and 10 for Indian Pines, Salinas and Enrique Reef respectively. As well as, the values of the PSNR obtained by the LSTM are greater than given for a CNN architecture.

On another side, the data of Enrique reef, Salinas and Indian Pines present a large set of homogeneous pixels over images. In future work, we propose to use images with a large set of heterogeneous pixels as well as evaluating the impact that this reconstruction approach has on image classification.

5. REFERENCES

[1] P. Ghamisi, B. Rasti, N. Yokoya, Q. Wang, B. Hofle, L. Bruzzone, F. Bovolo, M. Chi, K. Anders, R. Gloaguen, P. M. Atkinson, and J. A. Benediktsson, "Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 1, pp. 6–39, March 2019.

[2] Jixian Zhang, "Multi-source remote sensing data fusion: status and trends," *International Journal of Image and Data Fusion*, vol. 1, no. 1, pp. 5–24, 2010.

[3] Q. Wei, J. Bioucas-Dias, N. Dobigeon, and J. Tourneret, "Hyperspectral and multispectral image fusion based on a sparse representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 7, pp. 3658–3668, July 2015.

[4] B. Huang, H. Song, H. Cui, J. Peng, and Z. Xu, "Spatial and spectral image fusion using sparse matrix factorization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 3, pp. 1693–1704, March 2014.

[5] M. Q. Nguyen, P. M. Atkinson, and H. G. Lewis, "Superresolution mapping using a hopfield neural network with fused images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 3, pp. 736–749, March 2006.

[6] S. Zheng, W. Shi, J. Liu, and J. Tian, "Remote sensing image fusion using multiscale mapped ls-svm," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 5, pp. 1313–1322, May 2008.

[7] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, Feb 2016.

[8] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Multispectral and hyperspectral image fusion using a 3-d-convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 639–643, May 2017.

[9] M. Dalla Mura, G. Vivone, R. Restaino, P. Addesso, and J. Chanussot, "Global and local gram-schmidt methods for hyperspectral pansharpening," in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, July 2015, pp. 37–40.

[10] Sepp Hochreiter and Jrgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al., "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[12] M Fallah and A Azizi, "Quality assessment of image fusion techniques for multisensor high resolution satellite images (case study: Irs-p5 and irs-p6 satellite images)," vol. 38, 01 2010.