A NEW DEEP LEARNING METHOD FOR MULTISPECTRAL IMAGE TIME SERIES COMPLETION USING HYPERSPECTRAL DATA

C. T. Cissé[1], A. Alboody[1], M. Puigt[1], G. Roussel[1], V. Vantrepotte[2], C. Jamet[2], and T. K. 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

The massive development of remote sensing allowed many novel applications which bring new challenges. In particular, some applications such as marine observation require a good spatial, spectral, and temporal resolution. In order to tackle the last issue, spatio-temporal fusion of remote sensing data allows to complete a time series of multispectral images from, e.g., hyperspectral images. In this paper, we propose a new deep learning approach to that end. Our main contribution lies in the error completion task which allows to improve the completion performance. We show that our proposed method is able to produce high fidelity predictions with better quality indices than state-of-the-art methods on true images taken from the CIA / LGC database and Sentinel-2 / Sentinel-3 data.

Index Terms— Remote Sensing, Time-Series Completion, Spatio-Temporal Fusion, Deep Learning.

1. INTRODUCTION

The satellite observation of our planet knew significant instrumental advances for several decades, with consequent developments in terms of spatial resolution—e.g., in water remote sensing with high spatial resolution (10–60 m)—and in terms of spectral resolution (hyperspectral imagery). However, the Signal-to-Noise Ratio (SNR) of a Multispectral or Hyperspectral Imaging (MSI/HSI) sensor is proportional to the ratio between the sensor area and the number of observed spectral bands. Therefore, to maintain a constant SNR, increasing the number of spectral bands in an hyperspectral image implies a decrease in spatial resolution. As a consequence, our planet is currently observed by MSI systems having a very good spatial resolution but a low spectral resolution and by HSI systems having a very good spectral resolution but a low spatial resolution. Moreover, the sampling rate of remote sensing instruments may not necessarily be the same. However, for some applications—e.g., land use/cover classification [1] or change detection [2]—it may be necessary to observe an area with (i) a good spatial resolution, (ii) a good spectral resolution, and (iii) a good time resolution. Unfortunately, none of the actual satellite combine the three above properties. As an example, Sentinel-2 (S-2) has a high spatial resolution—ranging from 10 to 60 m—but only 13 spectral bands and a sampling rate around 5 days. On the contrary, Sentinel-3 (S-3) provides 21 spectral bands and a sampling rate around 1.4 days but with 300 m spatial resolution. While the fusion of MSI and HSI data acquired at the same time has been extensively investigated—see, e.g., [3, 4] for recent surveys and [5] for an application on S-2/S-3 data—the completion of MSI time series from HSI ones has been less investigated [6]. We here focus on the latter.

Popular state-of-the-art techniques are based on weighted filtering [7], weighted kriging [8], and/or weighted regression [9]. Other methods are based on unmixing, and extend multi-sharpening approaches to multi-temporal images [10]. Further techniques are learning dictionaries to perform the time series completion [11]. More recently, deep learning [12, 13] or hybrid techniques [14] were proposed.

In this paper, we aim to propose a new method for completing a time series of MS images from an HS one. Our proposed technique is based on a deep learning framework. The remainder of the paper reads as follows. In Section 2, we formally introduce the considered problem and our method. An experimental validation is provided in Section 3 while we conclude and discuss about future work in Section 4.

2. PROBLEM STATEMENT AND PROPOSED METHOD

In this section, we introduce in detail the considered problem. More specifically, we consider two time series of MS and HS images. We further assume that the time rate for HSI data is much lower than for MSI and that some HS images are acquired almost at the same time—i.e., the same days—as MS images. Without loss of generality, we consider in this paper that for one given MS image to estimate at Time t_2, there exist two pairs of MS and HS images acquired at t_1 and t_3, respectively, and one HS image acquired at t_2, as shown in Fig. 1. More specifically, we denote by $M(t_i)$ and $H(t_i)$ the MS and HS images sensed at Time t_i, respectively. As a

This work was partly funded by SFR "Campus de la Mer" and partly by CNES within the TOSCA "OSYNICO" project. Experiments presented in this paper were partly carried out using the CALCULCO computing platform, supported by SCoSI/ULCO.

consequence, we aim to derive $M(t_2)$ from $M(t_1)$, $M(t_3)$, $H(t_1)$, $H(t_2)$, and $H(t_3)$. Such an assumption was recently considered in the literature [14].

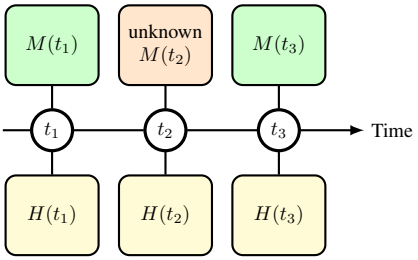

Fig. 1. Considered time sampling of MSI and HSI images.

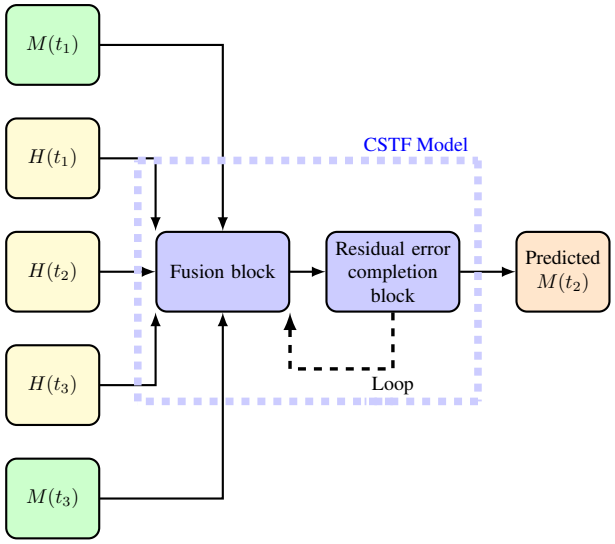

Fig. 2. Structure of the proposed approach.

In order to tackle this problem, we propose a deep learning method named Completion Spatio-Temporal Fusion (CSTF) which is composed of two main parts. The general structure of our proposed method is provided in Fig. 2. It consists of two blocks—i.e., a fusion and a residual completion error block—which are alternatingly run several times. Please note that the last time we run the latter, we run once more the former in order to provide the predicted image $M(t_2)$. The fusion block—whose structure is provided in Fig. 3—may be seen as a 5-image extension of the strategy proposed in [12] which used 3 images, i.e., 2 HS and 1 MS images. Indeed, we estimate the difference between pairs of HS images taken at adjacent times, that we then concatenate with available MS images in order to predict $M(t_2)$. The loss function considered in this paper is the mean squared error between the predicted and the target MS images at Time t_2. Our main contribution resides in the second block. Using the output of the fusion block, we get three MS images, i.e., $M(t_1)$, $M(t_3)$, and a first estimate of $M(t_2)$. These images

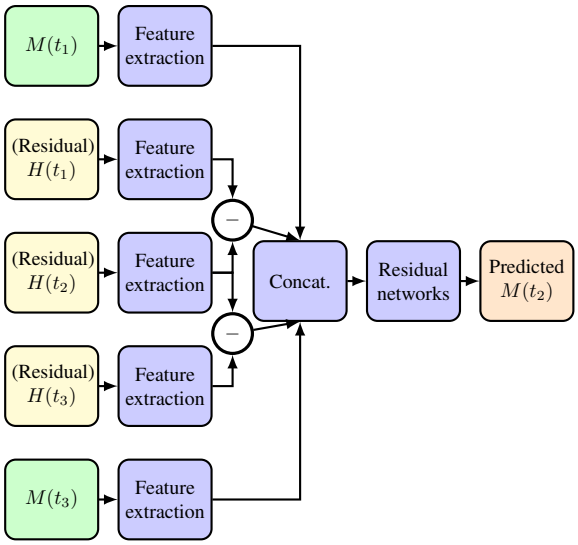

Fig. 3. Structure of the fusion block of CSTF.

can be compared $H(t_1)$, $H(t_3)$, and $H(t_2)$, respectively. More specifically, by spatially degrading the MS images and by considering the shared wavelengths with the HS images, we get quite similar images whose differences are used to improve the estimation of $M(t_2)$—both in terms of spatial and spectral information—in another fusion block. This procedure is repeated several times. The general structure of this residual completion error block is provided in Fig. 4.

From a mathematical point of view, the image $M(t_2)$ which is predicted during the fusion block is related to the available data through the following relationship:

$$\begin{aligned} M(t_2) =& h\Big(f_1\big(M(t_1)\big), f_2\big(M(t_3)\big), \\ & f_3(H(t_3)) - f_4(H(t_2)), \\ & f_5(H(t_1) - f_4(H(t_2))\Big), \end{aligned} \tag{1}$$

where $\forall i = 1, \ldots, 5$, $f_i(\cdot)$ is a feature extraction function—composed of 1 layer of convolution and 1 ReLu activation function—and $h(\cdot)$ is a function which concatenates the features of the different (differences of the) images and the residual networks. Please note that during the first pass in the fusion block, we directly process the HS images to predict $M(t_2)$. However, this predicted image is used as an input of the residual error completion block whose outputs are used as HS image inputs during the next passes of the fusion block.

The residual error completion block consists of three similar functions $g_i(\cdot)$ which are applied to $M(t_i)$ and $H(t_i)$ for $i = 1, \ldots, 3$, respectively. More precisely, for a given index $i \in \{1, 2, 3\}$, this function reads

$$H_r(t_i) = g_i(H(t_i), M(t_i)) = H(t_i) - f_i'\left(M(t_i)\right), \tag{2}$$

where $f_i'(\cdot)$ is the feature extraction function applied to the i-th MSI, and $H_r(t_i)$ is the residual HS image which is used as

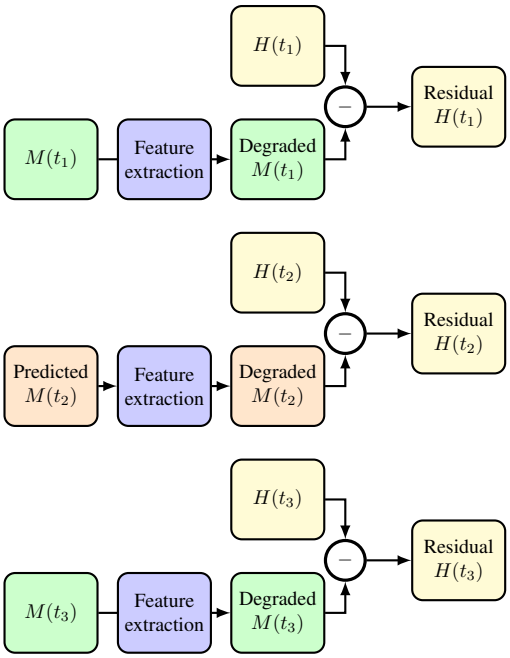

Fig. 4. Structure of the residual error completion block.

an input of the fusion block during the next step. Please note that each function $f_i'(\cdot)$ realizes a spatial degradation of the MS images for making them comparable with their respective HS ones but they also provide some specific processing, i.e., 3 convolutional layers, each with a ReLu activation function.

3. EXPERIMENTS

In this section, we investigate the ability of our proposed CSTF method to estimate a known MS image acquired at Time t_2 (not used in the estimation stage) from MS images acquired at Times t_1 and t_3 and HS images acquired at Times t_1, t_2, and t_3. For that purpose, we consider several databases and several methods. More precisely, we use two databases, i.e., the CIA/LGC databases [15] and some S-2 and S-3 time series. The former is a public database[1] which consists of MS and HS images with 25 m and 500 m spatial resolution and 6 spectral bands for both images. The latter consists of several images which are pre-processed using the method in [16] in order to remove atmospheric effects[2]. In this paper we use the 60 m spatial resolution of S-2 data. Both databases are then processed by several approaches, i.e., STARFM [7], DCSTFN [12], DMnet [6], DL-SDFM [14], and our proposed CSTF method. In particular, we extract patches from a time series of 24 pairs of images for the training. In order to apply several quantitative performance indices, we consider for each database 3 MS images to predict for which we know the ground truth, i.e., we have these images in the databases but we do not use them either for learning nor for predicting.

[1]The CIA and LGC databases may be found at `http://dx.doi.org/10.4225/08/5111AC0BF1229` and at `http://dx.doi.org/10.4225/08/5111AD2B7FEE6`, respectively.

[2]Such a pre-processing stage then allows the comparison of remote sensing data with *in situ* measurements, which is out of the scope of this paper.

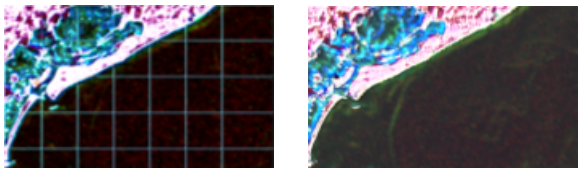

Fig. 5. Padding effect on the predicted MS images. Left: zero padding. Right: symmetrical padding.

Except DL-SDFM, the tested models like STARFM, DCSTFN, DMnet use a time series of 2 time indices t_1 and t_2 to perform the completion. For the sake of computation, all the deep learning models use patches for training. As a consequence, in order to predict a large image, it is necessary to cut it into 150×150 patches, to proceed to the prediction, and then to reconstruct the entire image from the patches. While it is the default strategy in several frameworks—e.g., Tensorflow—and is used in, e.g., DCSTFN, zero padding may provide some artifacts—visible on the left plot of Fig. 5—if the gaps between zero and the values at the border of the patch are not negligible. In order to provide a fair comparison, we replace it by symmetrical padding which removes this effect as shown on the right plot of Fig. 5.

In this paper, we set the parameters of our proposed method as explained in Tab. 1. The fusion block is run 6 times while the residual error completion block is run 5 times. In the fusion block, the size of the residual network is set to 10 convolutions.

Feature	Extraction parameters	Kernel	Filters
$f_1(M(t_1))$	Conv2D + Leaky ReLu	3×3	64
$f_2(M(t_2))$	Conv2D + Leaky ReLu	3×3	64
$f_5(H(t_1))$	Conv2D + Leaky ReLu	3×3	32
$f_4(H(t_2))$	Conv2D + Leaky ReLu	3×3	32
$f_3(H(t_3))$	Conv2D + Leaky ReLu	3×3	32
$f_1'(M(t_1))$	3 Conv2D + Leaky ReLu	1×1	64
$f_2'(M(t_2))$	Conv2D + Leaky ReLu	1×1	64

Table 1. Parameters of the functions in Eqs. (1) and (2).

In order to assess the performance of the tested methods, we use some classical quantitative performance measures for image quality assessment[3] , i.e., (i) the Peak Signal-to-Noise Ratio (PSNR) [17]—which is the ratio between the highest possible signal power and the noise power—(ii) the Spectral Angle Mapper (SAM) [18]—which is a pixelwise measure of the angle between the reference spectrum and the fused one.

[3]The code can be found at `https://github.com/andrewekhalel/sewar`.

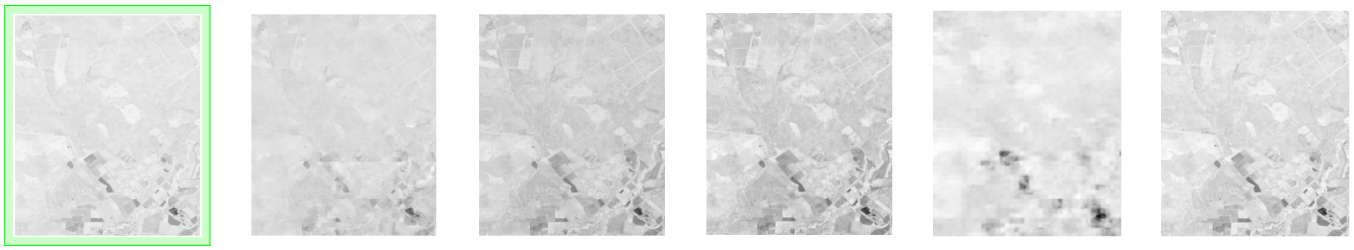

Fig. 6. Example of a predicted $M(t_2)$ image from CIA/LGC databases. From left to right: true image, outputs of STARFM, DCSTFN, DMnet, DL-SDFM, and our proposed method.

Method	PSNR	SAM	SSIM	SCC	UQI	PSNR	SAM	SSIM	SCC	UQI
Tests with 25 m spatial resolution on CIA/LGC						Tests with 60 m spatial resolution on S-2/S-3				
STARFM	24.2	0.6897	0.3974	0.1712	0.5357	17.7	0.9165	0.0531	**0.03155**	0.0692
DCSTFN	23.6	0.2187	0.7309	0.1211	0.7434	25.6	**0.4090**	0.1816	0.02138	**0.0791**
DMnet	24.0	0.2147	0.7421	0.2012	0.7529	19.2	0.4096	0.0555	0.02521	0.07382
DL-SDFM	21.4	0.2591	0.31460	0.03589	0.1720	30.1	0.4878	0.50374	-0.0023	0.0008
CSTF	**36.4**	**0.1277**	**0.92153**	**0.5225**	**0.9701**	**32.9**	0.7891	**0.6372**	0.03107	0.0748

Table 2. Performance of the tested methods on the considered databases.

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 Structural SIMilarity (SSIM) [17]—which measures the similarity between two given images and which appears more consistent than MSE in terms of visual perception. SSIM is bounded between 0 and 1 and SSIM values near 1 mean a high similarity.—(iv) the Spatial Correlation Coefficient (SCC) [19]—which is defined as the average correlation between pixels intensity—and (v) the Universal Image Quality Index (UIQI) which is designed by modeling any image distortion as a combination of three factors: loss of correlation, luminance distortion, and contrast distortion [20].

Please note that the performance indices from S-2/S-3 may be carefully considered. Indeed, these images may be cloudy, which generates two drawbacks. The first one is that S-2 and S-3 images are not taken at the same time, which implies that the cloud shapes and positions are not necessarily the same. In particular, our predicted $M(t_2)$ images have the same cloud shapes and positions than $H(t_2)$, which lower the prediction performance indices. Moreover, the pre-processing stage used in this database—to remove the atmospheric effects for a better water column observation—replaces the clouds by NaNs which are not taken into account in the computation of the performance values.

Table 2 provides the performance obtained for each database, for one unique spectral band at 60 m of spatial resolution. Our proposed method outperforms all the tested state-of-the-art methods for all the performance indices when applied to the CIA/LGC database. Please note that, as explained above, we used the symmetrical padding for all the methods, which allowed to improve the performance of state-of-the-art methods which are using zero padding. When applied to the S-2/S-3 database, our proposed approach provides the best performance for 2 performance indices, i.e., PSNR and SSIM. Except for the SAM index, the performance values reached for the other indices is close to the best ones. Let us stress again that our proposed method is almost always outperforming DL-SDFM which is also using 5 images for predicting $M(t_2)$ but which does not use our proposed residual error completion block.

As an example, we show in Fig. 6 some outputs obtained for one MSI image, observed at one wavelength, from the CIA/LGC database. One can notice that most of the spatial details are lost when STARFM and DL-SDFM are applied. The other approaches keep much more details but they appear slightly sharpener with our proposed CSTF method than with DCSTFN and DMnet. This shows the relevance of our work.

4. CONCLUSION

In this paper, we proposed a new deep learning method to perform time-series completion of MS images using HS ones. The structure of our network combines two blocks, i.e., a fusion and a residual error completion block, which are alternatingly called. Our main contribution lies in the former and the experiments conducted on real datasets show the relevance of our proposed approach. Our proposed residual error completion block can be applied to most of the state-of-the-art deep-learning-based time-series completion methods. In future work, we aim to better investigate the effects of its parameters as well as its application to other deep learning models, e.g., attention-based architectures. We would also like to deeply investigate the effects of clouds in the predicted image.

5. REFERENCES

[1] S. N. MohanRajan, A. Loganathan, and P. Manoharan, "Survey on land use/land cover (LU/LC) change analysis in remote sensing and GIS environment: Techniques and challenges," *Environmental Science and Pollution Research*, vol. 27, pp. 29900–29926, 2020.

[2] B. Sabouh, A. Alboody, M. N. Salah, and G. Hmeidan, "Change detection types of buildings in Aleppo citadel urban area during Syrian crisis using self-organizing maps neural networks and VHR QuickBird & Worldview-2 satellite images," in *Proc. IEEE IGARSS'21*, 2021.

[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] 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.

[5] 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," in *Proc. IEEE WHISPERS'21*, 2021.

[6] J. Li, Y. Li, L. He, J. Chen, and A. Plaza, "Spatio-temporal fusion for remote sensing data: An overview and new benchmark," *Science China Information Sciences*, vol. 63, no. 4, pp. 140301, 2020.

[7] F. Gao, J. Masek, M. Schwaller, and F. Hall, "On the blending of the landsat and modis surface reflectance: Predicting daily landsat surface reflectance," *IEEE Transactions on Geoscience and Remote sensing*, vol. 44, no. 8, pp. 2207–2218, 2006.

[8] J. Wang and B. Huang, "A rigorously-weighted spatiotemporal fusion model with uncertainty analysis," *Remote Sensing*, vol. 9, no. 10, pp. 990, 2017.

[9] Q. Wang and P. M. Atkinson, "Spatio-temporal fusion for daily sentinel-2 images," *Remote Sensing of Environment*, vol. 204, pp. 31–42, 2018.

[10] B. Zhukov, D. Oertel, F. Lanzl, and G. Reinhackel, "Unmixing-based multisensor multiresolution image fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 3, pp. 1212–1226, 1999.

[11] B. Huang and H. Song, "Spatiotemporal reflectance fusion via sparse representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 10, pp. 3707–3716, 2012.

[12] Z. Tan, P. Yue, L. Di, and J. Tang, "Deriving high spatiotemporal remote sensing images using deep convolutional network," *Remote Sensing*, vol. 10, no. 7, pp. 1066, 2018.

[13] W. Li, D. Cao, Y. Peng, and C. Yang, "MSNet: A multi-stream fusion network for remote sensing spatiotemporal fusion based on transformer and convolution," *Remote Sensing*, vol. 13, no. 18, pp. 3724, 2021.

[14] D. Jia, C. Song, C. Cheng, S. Shen, L. Ning, and C. Hui, "A novel deep learning-based spatiotemporal fusion method for combining satellite images with different resolutions using a two-stream convolutional neural network," *Remote Sensing*, vol. 12, no. 4, pp. 698, 2020.

[15] I. V. Emelyanova, T. R. McVicar, T. G. Van Niel, L. T. Li, and A. I. Van Dijk, "Assessing the accuracy of blending landsat–modis surface reflectances in two landscapes with contrasting spatial and temporal dynamics: A framework for algorithm selection," *Remote Sensing of Environment*, vol. 133, pp. 193–209, 2013.

[16] F. Steinmetz and D. Ramon, "Sentinel-2 MSI and sentinel-3 OLCI consistent ocean colour products using POLYMER," in *Proc. SPIE "Remote Sensing of the Open and Coastal Ocean and Inland Waters"*, 2018, vol. 10778.

[17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[18] R. H. Yuhas, A. F. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm," in *Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop*, 1992, vol. 1, pp. 147–149.

[19] J. Zhou, D. Civco, and J. Silander, "A wavelet transform method to merge landsat tm and spot panchromatic data," *International journal of remote sensing*, vol. 19, no. 4, pp. 743–757, 1998.

[20] Z. Wang and A. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002.