NTNU
Kunnskap for en bedre verden

Department of Computer Science

IMT4135 - Introduction to Research on Colour and Visual Computing

Review on Hyperspectral and Multispectral Image Fusion Methods

Author:
María José Rueda Montes

November, 2020

Contents

***Abstract*—high resolution hyperspectral image reconstruction (HR-HSI) by fusing hyperspectral images (LR-HSI) and high resolution multispectral images (HR-MSI), is a current research field whose results can be used in a large number of applications, such as remote sensing and computer vision tasks. The objective of this task is to combat the limitations of the sensors that capture hyperspectral images, since as the number of spectral bands captured increases, the spatial resolution decreases. In this way, it is intended to obtain a final image with a high spatial and spectral resolution. In this paper, a review of the most relevant methods in image fusion is carried out in order to address the structure of each method and give a general idea of the mechanism for obtaining HR-HSI.**

***Index Terms*—Hyperspectral and multispectral fusion, low resolution hyperspectral images (LR-HSI), high resolution multispectral images (HR-MSI), high resolution hyperspectral images (HR-HSI).**

1 Introduction

Hyperspectral images (HSI) are characterized by the hundreds of spectral bands obtained from a scene. These spectral bands include different wavelengths, covering ranges even outside the visible spectrum. The high spectral resolution of these type of images, allows good results to be obtained in applications for remote sensing and computer vision tasks (Dian *et al.*, 2018). The reason behind the successful results of hyperspectral imaging is related to its accurate identification of materials since each material has a different reflectance at each wavelength, capturing images at high spectral resolution and wide spectral range is an advanced discrimination between different materials in a scene (Dian *et al.*, 2020). However, due to the limitations of imaging sensors, long exposures capturing hyperspectral images are necessary, causing a spatial resolution deficiency due to the signal-to-noise-ratio (SNR) generated, which leads to the low spatial resolution of hyperspectral images (LR-HSI). On the other hand, imaging sensors can obtain an image with a higher spatial resolution but a worse spectral resolution, such as multispectral images (MSI), RGB images, or panchromatic images.

As a solution, and to improve the quality of hyperspectral images, HSI with high spectral resolution and MSI with high spatial resolution in the same scene can be combined. In this way, it is possible to obtain high resolution hyperspectral images (HR-HSI) in a process called fusion.

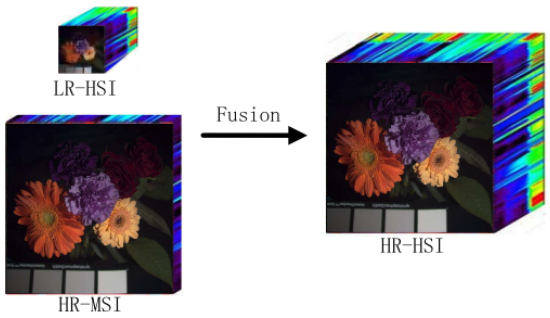

Figure 1: Hyperspectral image fusion using low resolution hyperspectral and high resolution multispectral images.

Source: (Dian *et al.*, 2018)

In this field, several satellite sensors such as IKONOS, GeoEye, OrbView, Landsat, SPOT, QuickBird, WorldView, and Pléiades, are able to detect and record the electromagnetic wave reflected by earth surfaces, simultaneously shooting a low resolution hyperspectral and high resolution panchromatic (PAN)/multispectral image. Unlike multispectral images, PAN images have a single band which includes red, green, and blue bands, the visible spectrum, each pixel contains the total

solar radiation intensity that is reflected by the objects. Because of the higher quantity of solar radiation contained in each pixel, panchromatic images are able to register brightness changes at smaller spatial sizes than multispectral images (Loncan *et al.*, 2015). In this way, ideally, the fused images should have the spatial resolution of high resolution images and preserve the spectral information in the images with a large number of spectral bands.

In order to obtain HR-HSI, there are different approaches in which each one handles the hyperspectral and multispectral images in a different way, being complex due to the fact that they include three dimensions, width, height, and spectral. In general aspects, four types of approaches can be differentiated: pan-sharpening, matrix factorization (MF), tensor representation (TR), and the most current, deep learning CNN methods.

This paper aims to review the methods used in the fusion of LR-HSI and HR-MSI for obtaining HR-HSI, in order to give a general idea of how each method works and some of the current research and contributions. Additionally, Section 3 discusses some of the most common databases and metrics used in experiments. Section 4 concludes with a short summary of the current challenges and future directions.

2 Fusion Methods

2.1 Pan-sharpening based HSI-MSI Fusion

Fusion process using low resolution Multispectral (MS) image and high resolution panchromatic (PAN) image is called pan-sharpening. The pan-sharpening has a very important and broad application in remote sensing, and they are divided into three main categories; CS-based methods, MRA-based methods, and VO-based methods.

The first method, Component substitution CS, is very common in panshapering, the most professional remote sensing software provide them. In the traditional methods, the MS image is transformed into a new suitable space. The spatial information component of the MS image is then replaced by the PAN image, and inverse transformation is carried out to reconstruct the fused image (Yuan *et al.*, 2018). Nevertheless, the current approach is easier, based on the substitution of the linear combination of spectral bands of the MS images, a single component, by the PAN image. In this manner, the high spatial structure information of the PAN image is extracted, calculated by the difference between the PAN image and the component that contains the spatial information in MS images. After this process, the spatial information extracted is introduced into the MS image. This procedure is notated as

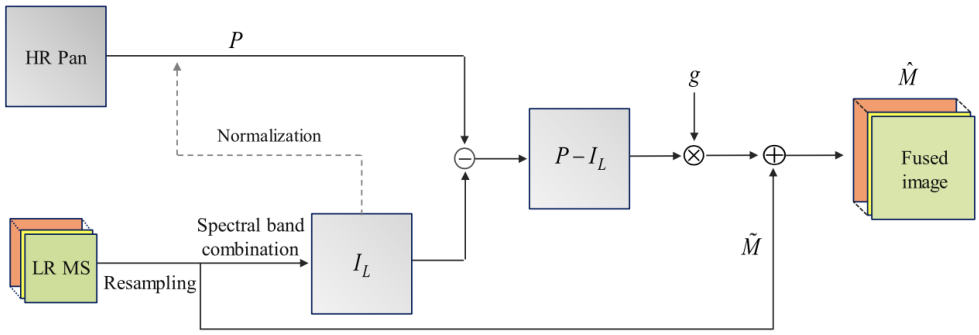

Figure 2: Steps used in component substitution pan-shapering fusion.

Source: (Meng *et al.*, 2019)

where $\hat{\mathbf{M}}$ represents the fused image, $\tilde{\mathbf{M}}$ the resampled MS image, $\mathbf{I}_L$ is the component which is replaced, $\mathbf{P}$ is the PAN image which is normalized with $\mathbf{I}_L$ to decrease the spectral distortion, and

g represents the injection weight (Meng *et al.*, 2019). The process can be represented as

$$\hat{\mathbf{M}} = \tilde{\mathbf{M}} + g(\mathbf{P} - \mathbf{I}_L) \tag{1}$$

Compared with the traditional CS-based methods, the multiresolution analysis MRA-based methods generally have better spectral information preservation (Yuan *et al.*, 2018). Briefly, this method consists in extracting the spatial information from the PAN image and applying a filter for generating the details, such as wavelet transform, Laplacian pyramid, contourlet transform, and couverlet transform methods. After this step, the spatial information extracted is injected into the upsampled MS image obtaining the final fused image (Loncan *et al.*, 2015)(Meng *et al.*, 2019). The representation of this method is

$$\hat{\mathbf{M}} = \tilde{\mathbf{M}} + g(\mathbf{P} - \mathbf{P}_L) \tag{2}$$

where $\mathbf{P}_L$ is the low-pass version of the PAN image. The main difference between this method and CS method is the way they extract the spatial information. In MRA methods it calculated the differences amidst PAN image $\mathbf{P}$ and its PAN low-pass version $\mathbf{P}_L$ to obtain the spatial structure information, as it is illustrated in Fig. 3. The $\mathbf{P}_L$ image is obtained after applying the correspondent kind of filter. The way of calculating the $\mathbf{P}_L$ image with the filters can be divided into two ways, calculation based on decimated filters and undecimated filters.

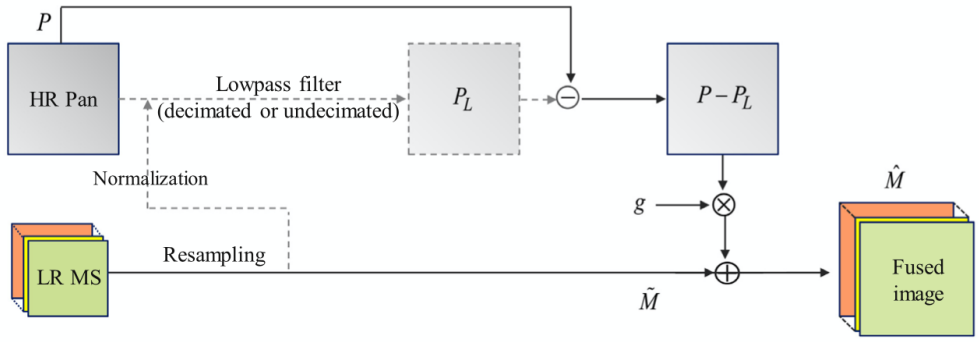

Figure 3: Multiresolution analysis (MRA) process.

Source: (Meng *et al.*, 2019)

Variational optimization (VO) methods represent an important category and consist mainly of the optimization of a variational model. In VO-based models, two main processes can be distinguished: the energy functional construction (A); and the optimization solution (B). Fig. 4 shows the steps until the model gets the final fused image (Meng *et al.*, 2019). In the energy functional construction, the aim is to make a construction of the optimal fusion energy functional. Within the process of energy functional construction, different methods can be found, such as image observation and sparse representation. Concerning the image observation methods, the model-based methods obtain the energy functional understanding of the entire fusion process as an ill-posed inverse optimization problem, in this way, the energy functional is generated based on the observation models between the ideal fused image and the degraded observations. As represented in Fig. 4, the LR-MSI image is supposedly obtained after applying blur, downsample, and noise to the ideal HR image that would be desired as a result of the fusion. On the other hand, the sparse-based theory methods represent the signal of the remote sensing image as a linear combination in an overcomplete dictionary, which is trained to improve the model (Meng *et al.*, 2019).

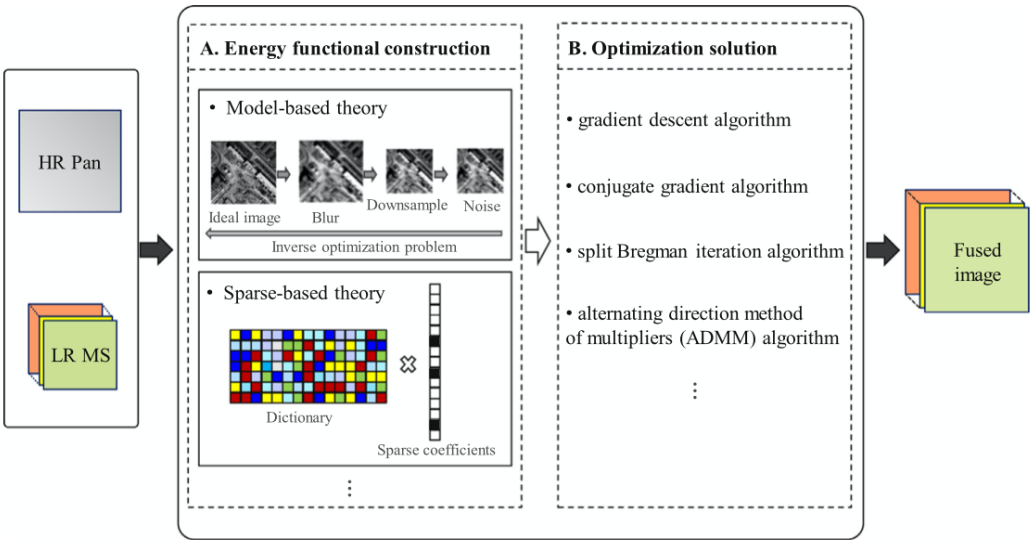

Figure 4: Variational optimization process in pan-sharpening methods.

Source: (Meng *et al.*, 2019)

The next block, the optimization solution, the fused image is achieved iteratively, operating with iterative optimization algorithms, such as gradient descent, conjugate, split Bergman iteration, or ADMM algorithms. After the last step, the most optimal solution of fused image is obtained, getting the final fused image (Meng *et al.*, 2019).

2.2 Matrix Factorization (MF) HR-HSI Fusion

MF-based methods unfold the tree dimensions of HR-HSI, considering this image is composed of a small number of spectral signatures (Dian *et al.*, 2018). In this way, the HR-HSI $\mathbf{X}_{(3)} \in \mathbb{R}^{S \times WH}$ can be approximated as a decomposition into a spectral basis $\mathbf{D}$ multiplied by coefficients $\mathbf{A}$ (Dian *et al.*, 2020), obtaining two two-dimensional matrices where the two dimensions denote the spatial locations and the band number (Zhang *et al.*, 2020).

$$\mathbf{X}_{(3)} = \mathbf{D}\mathbf{A}, \tag{3}$$

By sub-sampling HR-HSI is obtained HR-MSI

$$\mathbf{Y}_{(3)} = \mathbf{X}_{(3)}\mathbf{G}, \tag{4}$$

$$\mathbf{Z}_{(3)} = \mathbf{P}_3\mathbf{X}_{(3)}, \tag{5}$$

where $\mathbf{G}$ is the spatial dimensionality reduction operator and $\mathbf{P}$ the spectral dimensionality reduction operator, while the matrices $\mathbf{Y}_{(3)} \in \mathbb{R}^{S \times wh}$ and $\mathbf{Z}_{(3)} \in \mathbb{R}^{s \times WH}$ are the spectral mode (3-mode) unfolding matrices of $\mathcal{Y}$ and $\mathcal{Z}$, respectively (Li *et al.*, 2018). Depending on the way to model the spectral basis, MF can be classified as sparse representation methods and low-rank methods. The sparse representation-based method represents each HR-HSI pixel as a linear combination of spectral signatures. The spectral basis $\mathbf{D}$ is considered as the over-complete dictionary. In this dictionary, each atom means the spectral vector of materials in the image (Dian *et al.*, 2020). In order to learn dictionaries, sparse dictionaries learning algorithms are applied to LR-HSI, while the coefficients are estimated by applying sparse coding algorithms. Otherwise, the purpose of low-rank-based methods is to represent the spectral signatures with low dimensions, where the spectral basis $\mathbf{D}$ is a low-rank matrix. $\mathbf{D}$ matrix is learned from LR-HSI. The reduction of the dimensions has as a consequence an improvement in the fusion speed (Dian *et al.*, 2020).

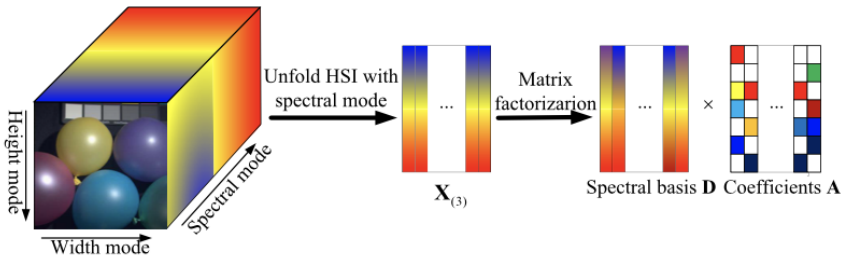

Figure 5: Matrix factorization representation.

Source: (Dian *et al.*, 2020)

There are different MF models based on how they solve the fusion problem, and obtain the spectral basis and coefficients estimation. For example, Yokoya *et al.* introduce the nonnegative matrix factorization, where the spectral information is acquired from the HR-MSI and the spatial information from LR-HSI, the last one integrated for generating the HR-HSI. Another proposal for Akhtar *et al.* design a simultaneous greedy pursuit algorithm that applies to each pixel patch to obtain the sparse coefficients, thus taking advantage of the similarity of nearby pixels to each other. Dong *et al.* explain a sparsity-based hyperspectral image super-resolution method where the over-complete dictionary is learned from the LR-HSI through a nonnegative dictionary-learning and the coefficients are estimated using sparse priors (Zhang *et al.*, 2020)(Dian *et al.*, 2018). Dian *et al.* proposed SSSR method, witch express each pixel as a linear combination of similar pixels, and calculate the coefficients with sparse priors (Dian *et al.*, 2020).

Taking into account the structure of the MF methods, it is difficult to study precisely the spatial and spectral correlations since these methods use two different matrices to compose the 3D HR-HSI. Tensor-based methods are proposed as a solution to this problem (Zhou *et al.*, 2019).

2.3 Tensor representation (TR) HR-HSI Fusion

(Dian, *et al.*, 2017) propose the tensor factorization-based method. Unlike MF methods, HR-HSI is interpreted as a core tensor multiplied by the dictionaries (factor matrices) of the three modes, width, height, and spectral. The model is shown in Fig. 6. This method facilitates working with HSI since this type of image have three dimensions and can be translated into a three-dimensional tensor. This type of structure manages to obtain good results that are lately applied in reconnaissance, unmixing, data restoration, visual tracking, object detection, and other fields.

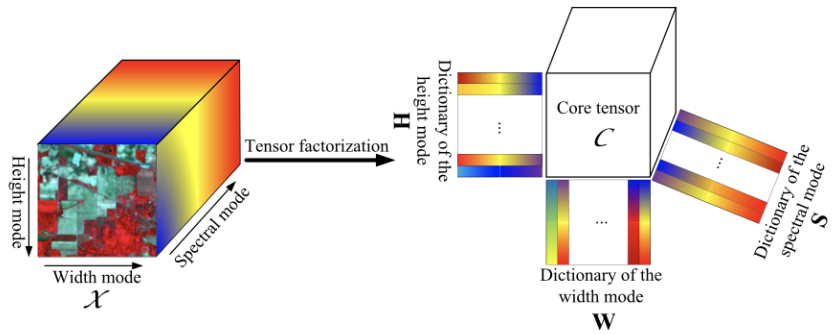

Figure 6: Tensor decomposition of HR-HSI.

Source: (Li *et al.*, 2018)

An important method in tensor representation is Tucker decomposition, where the 3D tensor is

factored in a main tensor (Core tensor) multiplied by the factor matrix of each dimension. This decomposition allows saving the information of each dimension in a tensor, obtaining the three tensors related by the main tensor (Li *et al.*, 2018).

The tensor factorization of the HR-HSI can be formulated as follows

$$\mathcal{X} = \mathcal{C} \times_1 \mathbf{W} \times_2 \mathbf{H} \times_3 \mathbf{S}, \tag{6}$$

where the matrices $\mathbf{W} \in \mathbb{R}^{W \times n_w}$, $\mathbf{H} \in \mathbb{R}^{H \times n_h}$ and $\mathbf{S} \in \mathbb{R}^{S \times n_s}$ represent the dictionaries of with, height and spectral modes respectively, with n_w, n_h and n_s atoms. $\mathcal{C} \in \mathbb{R}^{n_w \times n_h \times n_s}$ is the core tensor which include the coefficients of $\mathcal{X}$ over the dictionaries. HR-HSI spectral vectors can be approximated in a low dimensional subspace, being able to find a small number of atoms n_s to represent the spectral dictionary $\mathbf{S}$, while the characteristic of spatial-similarity of HR-HSI allows to make a sparse representation of the dictionaries $\mathbf{W}$ and $\mathbf{H}$. To obtain HR-HSI, the two aforementioned characteristics are intended to be achieved, which can be obtained by reducing the dimensions of core tensors. In this line, considering the point spread function (PSF) of the hyperspectral sensor and that the matrices of width and height mode can be separated, LR-HSI $\mathcal{Y}$ can be decomposed into

$$\mathcal{Y} = \mathcal{X} \times_1 \mathbf{P}_1 \times_2 \mathbf{P}_2, \tag{7}$$

where $\mathbf{P}_1 \in \mathbb{R}^{w \times W}$ and $\mathbf{P}_2 \in \mathbb{R}^{h \times H}$ denotes the downsampling matrices on the width and height modes, respectively. Applying the separability assumption and using the equations (6) and (7), $\mathcal{Y}$ can be rewritten as

$$\mathcal{Y} = \mathcal{C} \times_1 (\mathbf{P}_1\mathbf{W}) \times_2 (\mathbf{P}_2\mathbf{H}) \times_3 \mathbf{S} = \mathcal{C} \times_1 \mathbf{W}^* \times_2 \mathbf{H}^* \times_3 \mathbf{S}, \tag{8}$$

where $\mathbf{W}^* = \mathbf{P}_1\mathbf{W} \in \mathbb{R}^{w \times n_w}$ and $\mathbf{H}^* = \mathbf{P}_2\mathbf{H} \in \mathbb{R}^{h \times n_h}$ are the downsampled dictionaries along the width and height modes, respectively. Additionally, the separability assumption also means the spatial subsampling matrix $\mathbf{M}$ can be decomposed with respect the two spatial modes, obtaining

$$\mathbf{M} = (\mathbf{P}_2 \otimes \mathbf{P}_1)^T, \tag{9}$$

The HR-MSI $\mathcal{Z}$ is the spectral downsampling of $\mathcal{X}$, as $\mathcal{Y}$ is the spatial. $\mathcal{Z}$ can be written as

$$\mathcal{Z} = \mathcal{X} \times_3 \mathbf{P}_3, \tag{10}$$

where $\mathbf{P}_3 \in \mathbb{R}^{s \times S}$ represent the downsampling matrix of spectral mode. As with LR-HSI in (8), it can be obtained from the spectral dictionary as

$$\mathcal{Z} = \mathcal{C} \times_1 \mathbf{W} \times_2 \mathbf{H} \times_3 \mathbf{S}^*, \tag{11}$$

where $\mathbf{S}^* = \mathbf{P}_3\mathbf{S} \in \mathbb{R}^{s \times n_s}$ is the downsampled spectral dictionary.

In order to reconstruct the final HR-HSI, the aim is to calculate the dictionaries $\mathbf{W}$, $\mathbf{H}$ and $\mathbf{S}$ its corresponding core tensor $\mathcal{C}$. (Dian *et al.*, 2017) propose a HSI a novel NLSTF method, which combines the sparse tensor factorization and the non-local means approach. This method considers separating the HR-HSI into cubes, where each cube learns the spatial dictionaries $\mathbf{W}$ and $\mathbf{H}$ from HR-MSI, and the spectral dictionary $\mathbf{S}$ from LR-HSI. Finally, the core tensor is estimated by sparse prior. Later, (Li *et al.*, 2018) use sparse tensor factorization based on Tucker decomposition CSTF. This technique estimates the core, and the spatial and spectral dictionaries as a sparse tensor decomposition of the HR-MSI and LR-HSI. Furthermore, Canonical polyadic CP decomposition, a type of Tucker decomposition, divide the N-dimension tensor in N-factor matrices. Based on CP, (Xu *et al.*, 2019) propose a non-local CP decomposition for HR-HSI fusion, since the LR-HSI and HR-MSI share the same factor matrices in CP decomposition. Furthermore (Dian, Li, and Fang, 2019) propose a low tensor-train rank LTTR prior which learn the relation between the spectral, spatial, and non-local modes of the HR-HSI cubes. The similar cubes are grouped in a four-dimensional tensor, and then the ADMMs algorithm is used to solve the optimization problem (Zhang *et al.*, 2020)(Dian *et al.*, 2020).

2.4 Deep Learning (CNN) HR-HSI Fusion

In the last decade, methods based on Deep Learning, especially convolutional neural networks CNN, have achieved pioneering results in a wide variety of fields, especially in applications that include image data: such as image classification, computer vision, image processing, etc. In the field of LR-HSI and HR-MSI fusion, deep learning approaches is achieving excellent results as well. the mechanism of CNN is based on the neural functioning of the visual cortex, in this way the CNNs are structured in different layers which are in charge of extracting the features of the images in order to later train the model and in this way update it based on the task to be achieved, be it classification, detection, prediction, or as in this case, the fusion of LR and HR images. Throughout the convolutional layer, a numbers array of a given size is slipped or convolved trough the image, in order to obtain edge information. This matrix is called a filter or kernel. As the filter moves over a certain region of the image (receptive field), the filter values (weights or parameters) and the values of the receptive field region are multiplied, obtaining a new numbers array is added to give the output value that will serve as input to the corresponding next layer filter. For LR and HR fusion, the filters are focused on extracting the spatial and spectral information from the images, generating the corresponding edge maps. After the first convolutional layer, the generated feature map is used as input for the second convolutional layer, and so on. Therefore, and since the filters of the next layer only take the information from their receptive field, each layer describes the locations of the image in which the features are found. In HR-HSI fusion, in order to be able to make a spatial and spectral reconstruction, the loss function updates the weigh of each filter. In addition, other layers are incorporated during the process such as nonlinear ReLU function as activation layer to improve the processing.

The CNN models for LR-HSI and HR-MSI fusion can be divided into two types: input-level fusion and feature level fusion. For input-level fusion, the LR-HSI and HR-MSI images are fused before passing through the neural network to obtain the final fused HR-HSI. Normally, before the fusing, interpolation is applied to the LR-HSI to obtain the same size as the HR-MSI. Works of (Masi *et al.*, 2016) or (Palsson *et al.*, 2017) use the input-level fusion, where the preliminary fused LR-HSI and HR-MSI are used as input for a super-resolution CNN, SRCNN, and PCA prior for reducing the dimensions of the fusion. (Diang *et al.*, 2018) present a DHSIS method to reconstruct HR-HSI, where they map the first fused HR-HSI to the reference HR-HSI using the priors learned from a deep CNN-based residual learning to regularize the fusion.

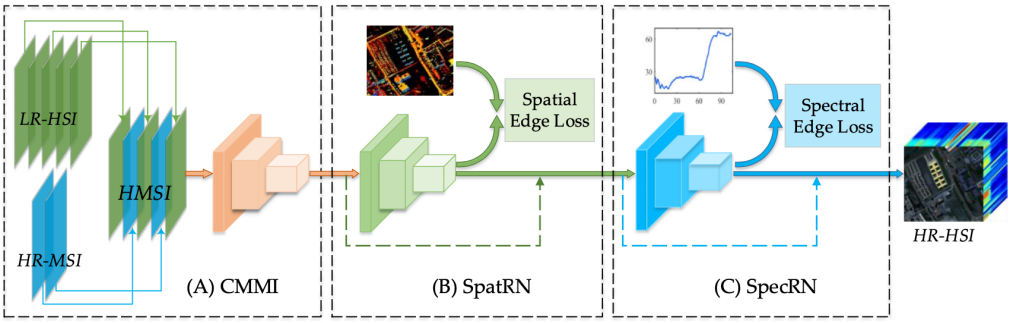

Figure 7: Example of hyperspectral image input-level fusion model, SSRNET.

Source: (Zhang *et al.*, 2020)

In feature-level fusion approaches, first the spatial features are extracted from HR-MSI and the spectral features from LR-HSI. Both features are fused in order to reconstruct the HR-HSI. (Shao *et al.*, 2018) proposed a model with two branches network to extract the HR-MSI and LR-HSI information separately. (Yang *et al.*, 2018) use two branches for extracting the spectral features of each pixel in LR-HSI, and its correspondent spatial neighborhood in HR-MSI, to take advantage of the spatial correlation. Then they fuse the information in a fully connected layer. (Hang *et al.*, 2019) a multi-scale CNN system, which reduces the HR images and increases the feature sizes of LR in a gradual way. In addition, they add the multi-level cost functions in the train process for solving the vanish gradient problem.

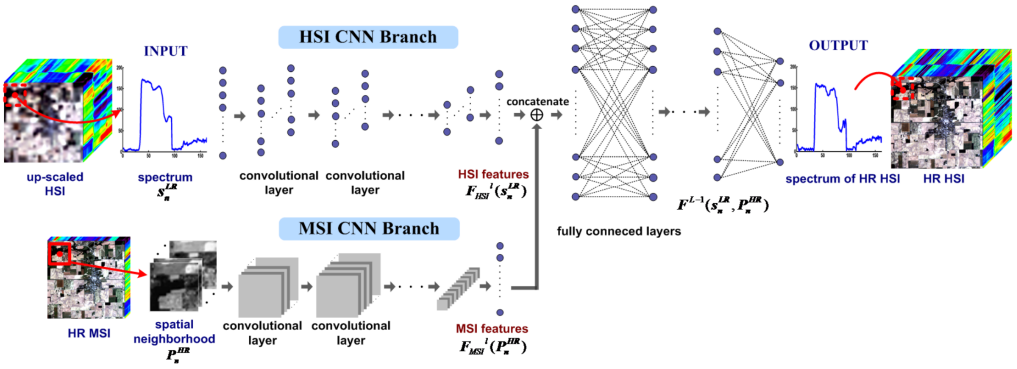

Figure 8: Example of hyperspectral image feature level fusion model.

Source: (Yang *et al.*, 2018)

3 Experiments

In the set of experiments it is important to have different databases and different measurement metrics. In this way, it is intended to evaluate the different methods in the most objective way, and to be able to be compared with each other.

3.1 Data Sets

Most of the databases that are going to be exposed below are available as an open resource, in addition the images have their corresponding ground truth.

1) Botswana

This database was captured in 2001-2004 by the Hyperion sensor of the NASA EO-1 satellite over the Okavango Delta. The data consist of 147 x 256 pixels with a spatial resolution of 30 m. Spectral bands cover wavelengths from 400 to 2500 nm. Removing the uncalibrated and noisy bands of water absorption features 145 bands left.

2) Pavia University

Database obtained in 2003 by ROSIS-3 optical airborne sensor over the area of the University of Pavia, Italy. This dataset has 610 x 610 pixels and 1.3 m resolution. The bands are 115 with a spectral range of 430 to 860 nm within an interval of 10 nm.

3) Pavia Center

The Pavia Center database was collected with the same optical sensor used to capture the images from Pavia University, thus having the same spatial resolution, 1.3 m. However, in this case each band has 1096 x 1096 pixels.

4) Washington DC Mall

The Washington DC Mall (WDCM) data set was obtained in 1995 by the hyperspectral digital imagery collection experiment (HYDICE) sensor, over the National Mall in Washington, DC. These dataset consist of 191 bands from 400 to 2500 nm with a spatial resolution of 2.5 m and 1280 x 307 dimensions. The number of bands can be reduced in 191 removing the bands covering the region of water absorption.

5) Indian Pines

This data was taken by the AVIRIS sensor over the Indian Pines test site in North-western, Indiana. The database has 145 x 145 pixels with 224 bands from 400 to 2500 as spectral range.

Indian Pines dataset contains scenes of vegetation and natural landscapes, in addition to scenes of railways, highways and some buildings that are not very crowded. By deleting the bands covering the region of water absorption, there are 200 bands remaining.

6) Urban

Urban dataset was collected in 1995 over Copperas Cove, TX, USA. There are 307 x 307 pixels with 2 m as spatial resolution. Furthermore, this dataset is composed of 210 bands in total in a range from 400 to 2500 nm with an interval of 10 nm. Urban is made up of captures of buildings, architectural structures, or urban landscapes. When removed bands of dense water vapor and atmospheric, are obtained 162 bands.

3.2 Evaluation Metrics

1) Peak Signal-to-Noise Ratio (PSNR)

The peak SNR (PSNR) is a very popular quality metric, used to evaluate the spatial quality of the bands in the reconstructed HR-HSI. This metric measures the similarities between the fused image and the reference image.

$$PSNR = 10 \log_{10} \left(\frac{\max(\mathbf{R}_k)^2}{\frac{1}{HW} \parallel \mathbf{R}_k - \mathbf{Z}_k \parallel_2^2} \right), \tag{12}$$

where R and Z are the *kth* band of the reference and estimated fused image, respectively. The final result is the average of all the bands, where a higher result indicates a better spatial quality of the fused image.

2) Erreur Relative Globale Adimensionelle de Synthèse (ERGAS)

The ERGAS measures the quality of the fused image in a global statistical way, where the ideal value would be 0.

$$ERGAS = \frac{100}{r} \sqrt{\frac{1}{L} \sum_{k=1}^{L} \frac{\parallel \mathbf{R}_k - \mathbf{Z}_k \parallel_2^2}{\mu^2(\mathbf{R}_k)}}, \tag{13}$$

where r is the ratio of the spatial resolution from HR-MSI to LR-HSI. R_k and Z_k denotes the *kth* bands of the reference and fused image, accordingly. Moreover, $\mu(R_k)$ represents the mean value of the *kth* band in the reference image.

3) Spectral Angle Mapper (SAM)

This metric evaluates the spectral information preservation at each pixel, which is important to estimate the spectral distortion.

$$SAM = \arccos \left(\frac{\langle \mathbf{R}(i,j), \mathbf{Z}(i,j) \rangle}{\parallel \mathbf{R}(i,j) \parallel_2 \parallel \mathbf{Z}(i,j) \parallel_2} \right), \tag{14}$$

where $\mathbf{R}(i,j)$ and $\mathbf{Z}(i,j)$ represent the spectral vectors at the spatial pixel position (i,j) in the reference and estimated fused image respectively. In addition, $\langle \mathbf{R}(i,j), \mathbf{Z}(i,j) \rangle$ is the two vector inner product. The final result is obtained by averaging the SAM metric for all pixels. A better spatial quality is obtained for SAM values close to 0.

4) Root-Mean-Squared Error (RMSE)

The RMSE measures the difference between the estimated and reference images, to compare the

prediction errors.

$$RMSE = \sqrt{\frac{\sum_{k=1}^{L}\sum_{i=1}^{H}\sum_{j=1}^{W}(\mathbf{R}_k(i,j) - \mathbf{Z}_k(i,j))^2}{HWL}}, \tag{15}$$

The best results are obtained with smaller values for the RMSE.

5) Structural Similarity Index (SSIM)

SSIM metric is applied to each band of the estimated and reference images, measuring the structural similarities between the two images. The average of all the bands is the final result. A higher result means better image quality.

4 Challenges and New Guideline

The LR-HSI and HR-MSI fusion models can be divided into two blocks: traditional methods, such as pan-sharpening, matrix factorization MF or tensor representation TR; and new methods, which include CNN techniques.

A significant body of research exists on pan-sharpening, matrix factorization, and tensor representation-based methods. Multiple techniques have been developed for each type of approach, however, unresolved issues remain. In pan-sharpening methods, the spectral difference between the MS and PAN images produces distortion problems, in addition, another frequent problem is the misalignments that produce spatial artifacts in the fused image, especially in multiresolution analysis methods. So future approaches should be oriented at fixing the different spectral responses between multispectral and PAN images, as well as exploring methods with more resistance to misalignments. On the other hand, MF-based methods depend to a great extent on the selection of parameters, which are also difficult to achieve. Future approaches to these methods are focused on exploring techniques that improve optimization and reduce the computational cost of these systems. TR-based methods solve some of the problems that arose in MF-based methods, such as the improvement in the study of spatial and spectral correlation, although they still have the same optimization problem, so the new approaches consider improving the processing by incorporating different optimization algorithms.

Summing up the traditional methods, these are dependent on the choice of features, as well as being difficult to extract. In order to obtain the features, complex processing of the images with a great computational cost is required. In addition, the choice of the parameters is a challenge. As a new approach and in order to solve the complexity of the feature extraction, new methods that use deep learning CNN are proposed. Although these models are obtaining pioneering results and the success outcomes indicate the future is focused on the use of these systems, they also encounter some drawbacks. These systems need large databases to give good results, which in view of the amount of hyperspectral and multispectral images available is considerably less than RGB images, is a challenge.

5 Conclusion

This paper shows a review of fusion methods to achieve high-resolution hyperspectral images HR-HSI, dividing the methods into four different approaches. In addition, the most relevant databases and metrics used in the set of experiments are mentioned. The project concludes with a description of the problems arising from each method and the new research directions.

6 References

Dian, R., Fang, L., Li, S. (2017). Hyperspectral image super-resolution via non-local sparse tensor factorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5344-5353).

Dian, R., Li, S., Fang, L. (2019). Learning a low tensor-train rank representation for hyperspectral image super-resolution. *IEEE transactions on neural networks and learning systems, 30*(9), 2672-2683.

Dian, R., Li, S., Fang, L., Lu, T., Bioucas-Dias, J. M. (2019). Nonlocal sparse tensor factorization for semiblind hyperspectral and multispectral image fusion. *IEEE Transactions on Cybernetics*

Dian, R., Li, S., Guo, A., Fang, L. (2018). Deep hyperspectral image sharpening. *IEEE transactions on neural networks and learning systems*, (99), 1-11.

Dian, R., Li, S., Sun, B., Guo, A. (2020). Recent Advances and New Guidelines on Hyperspectral and Multispectral Image Fusion. *arXiv preprint arXiv:2008.03426.*

Fang, L., Zhuo, H., Li, S. (2018). Super-resolution of hyperspectral image via superpixel-based sparse representation. *Neurocomputing*, 273, 171-177.

Han, X. H., Zheng, Y., Chen, Y. W. (2019). Multi-Level and Multi-Scale Spatial and Spectral Fusion CNN for Hyperspectral Image Super-Resolution. *In Proceedings of the IEEE International Conference on Computer Vision Workshops* (pp. 0-0).

Kanatsoulis, C. I., Fu, X., Sidiropoulos, N. D., Ma, W. K. (2018, October). Hyperspectral super-resolution: Combining low rank tensor and matrix structure. In *2018 25th IEEE International Conference on Image Processing (ICIP)* (pp. 3318-3322). IEEE.

Kilmer, M. E., Braman, K., Hao, N., Hoover, R. C. (2013). Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging. *SIAM Journal on Matrix Analysis and Applications, 34*(1), 148-172.

Li, S., Dian, R., Fang, L., Bioucas-Dias, J. M. (2018). Fusing hyperspectral and multispectral images via coupled sparse tensor factorization. *IEEE Transactions on Image Processing, 27*(8), 4118-4130.

Liu, X., Zhai, D., Bai, Y., Ji, X., Gao, W. (2019). Contrast Enhancement via Dual Graph Total Variation-Based Image Decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8), 2463-2476.

Loncan, L., De Almeida, L. B., Bioucas-Dias, J. M., Briottet, X., Chanussot, J., Dobigeon, N., ... Tourneret, J. Y. (2015). Hyperspectral pansharpening: A review. *IEEE Geoscience and remote sensing magazine*, 3(3), 27-46.

Masi, G., Cozzolino, D., Verdoliva, L., Scarpa, G. (2016). Pansharpening by convolutional neural networks. *Remote Sensing, 8*(7), 594.

Meng, X., Shen, H., Li, H., Zhang, L., Fu, R. (2019). Review of the pansharpening methods for remote sensing images based on the idea of meta-analysis: Practical discussion and challenges. *Information Fusion*, 46, 102-113.

Oseledets, I. V. (2011). Tensor-train decomposition. *SIAM Journal on Scientific Computing, 33*(5), 2295-2317.

Palsson, F., Sveinsson, J. R., Ulfarsson, M. O. (2017). Multispectral and hyperspectral image fusion using a 3-D-convolutional neural network. *IEEE Geoscience and Remote Sensing Letters, 14*(5), 639-643.

Qu, J., Lei, J., Li, Y., Dong, W., Zeng, Z., Chen, D. (2018). Structure tensor-based algorithm for hyperspectral and panchromatic images fusion. *Remote Sensing, 10*(3), 373.

Shao, Z., Cai, J. (2018). Remote sensing image fusion with deep convolutional neural network. *IEEE journal of selected topics in applied earth observations and remote sensing, 11*(5), 1656-1669.

Xie, Q., Zhou, M., Zhao, Q., Meng, D., Zuo, W., Xu, Z. (2019). Multispectral and hyperspectral image fusion by MS/HS fusion net. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1585-1594)

Xu, Y., Wu, Z., Chanussot, J., Comon, P., Wei, Z. (2019). Nonlocal coupled tensor cp decomposition for hyperspectral and multispectral image fusion. *IEEE Transactions on Geoscience and Remote Sensing, 58*(1), 348-362.

Yang, J., Zhao, Y. Q., Chan, J. C. W. (2018). Hyperspectral and multispectral image fusion via deep two-branches convolutional neural network. *Remote Sensing, 10*(5), 800.

Yuan, Q., Wei, Y., Meng, X., Shen, H., Zhang, L. (2018). A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(3), 978-989.

Zhang, X., Huang, W., Wang, Q., Li, X. (2020). SSR-NET: Spatial-Spectral Reconstruction Network for Hyperspectral and Multispectral Image Fusion. *IEEE Transactions on Geoscience and Remote Sensing*.

Zhou, F., Hang, R., Liu, Q., Yuan, X. (2019). Pyramid fully convolutional network for hyperspectral and multispectral image fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12*(5), 1549-1558.

.

Appendix

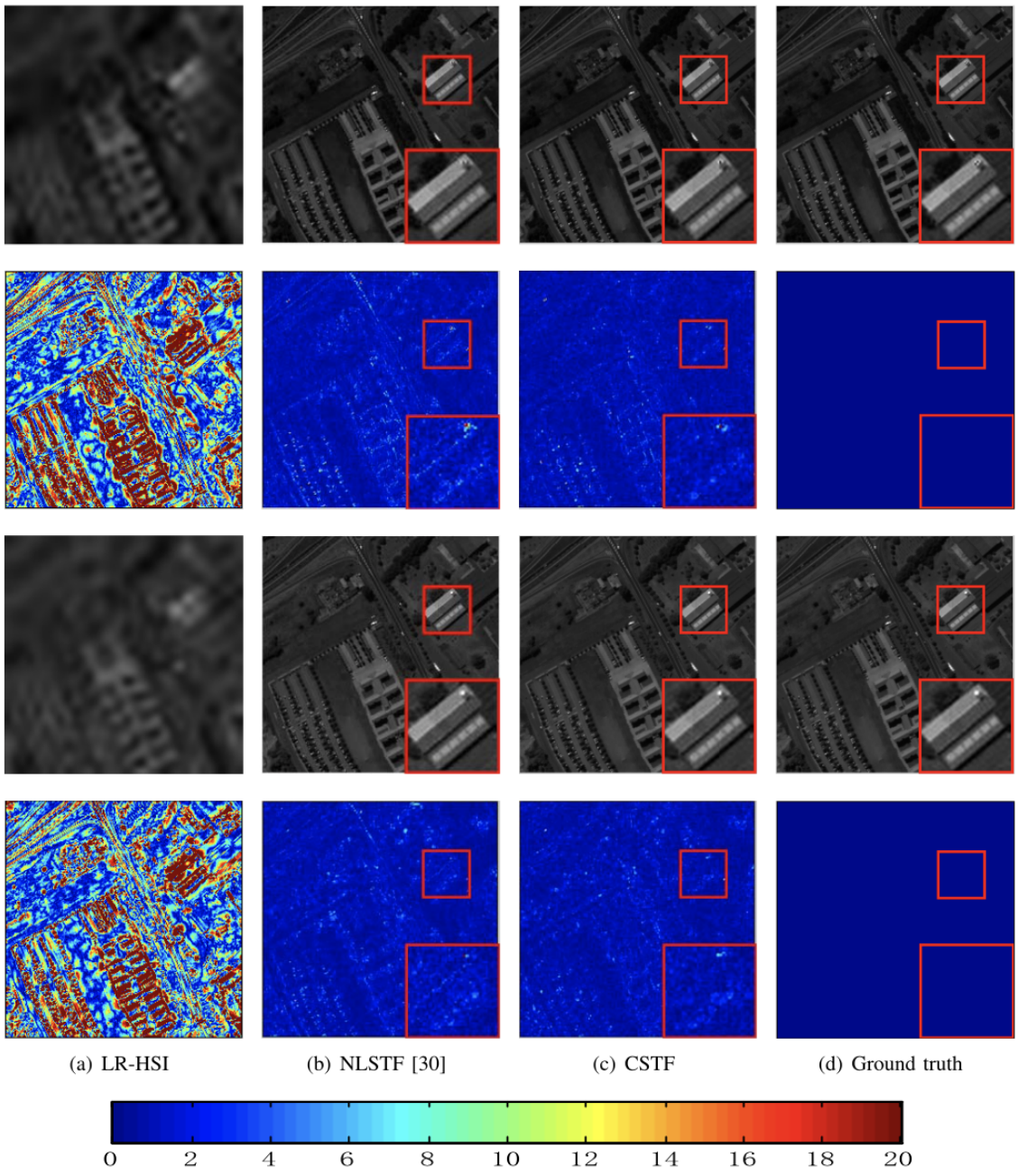

Figure 9: Example of reconstructed images using Tensor Factorization method and corresponding error images of Pavia University. The first and second rows show the reconstructed images for the 40th band corresponding error images, respectively. The third and forth rows are the reconstructed images for the 60th band and corresponding error images, respectively. (a) LR-HSI; (b) the NLSTF method; (c) the proposed CSTF method; (d) Ground truth.

Source: (Li *et al.*, 2018)

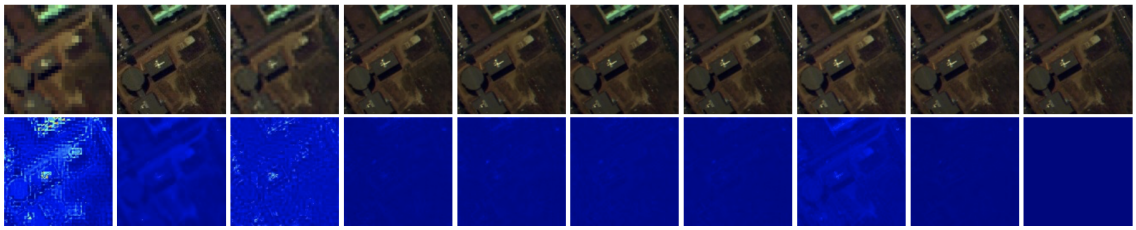

Figure 10: Example of Deep Learning fusion results of SSRNET in Pavia University dataset, where 'ConSSF' and 'GT' respectively reprensent the ConSSFCNN and the ground-truth image. The first row shows the RGB images, 67-29-1 bands, of the estimated HR-HSI, and the second row shows the difference images between the estimated and the reference RGB images.

Source: (Zhang *et al.*, 2020)

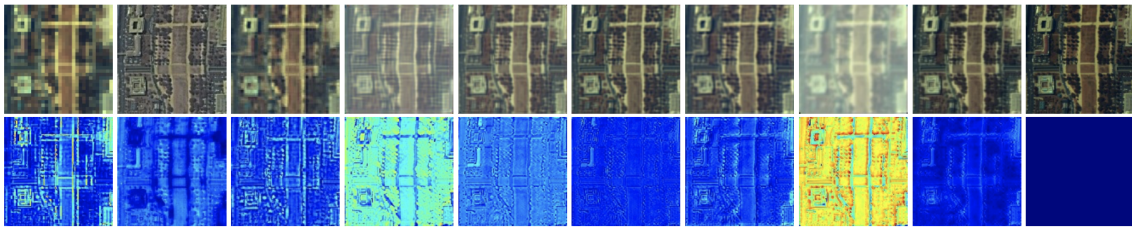

Figure 11: Example of Deep Learning fusion results of SSRNET in Washington DC Mall dataset, where 'ConSSF' and 'GT' respectively reprensent the ConSSFCNN and the ground-truth image. The first row shows the RGB images, 55-35-11 bands, of the estimated HR-HSI, and the second row shows the difference images between the estimated and the reference RGB images.

Source: (Zhang *et al.*, 2020)

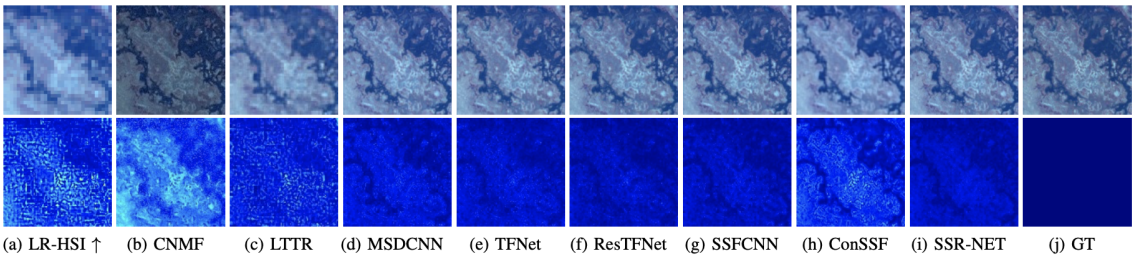

Figure 12: Example of Deep Learning fusion results of SSRNET in Botswana dataset, where 'ConSSF' and 'GT' respectively reprensent the ConSSFCNN and the ground-truth image. The first row shows the RGB images, 48-15-4 bands, of the estimated HR-HSI, and the second row shows the difference images between the estimated and the reference RGB images.

Source: (Zhang *et al.*, 2020)