Deep unfolding for hypersharpening using a high-frequency injection module

Jamila Mifdal [1] Marc Tomás-Cruz [2] Alessandro Sebastianelli[1] Bartomeu Coll[2]
Joan Duran [2]
[1] Φ-lab, European Space Agency, ESRIN 00044 Frascati, Italy
[2]DMI & IAC3, Universitat de les Illes Balears, Cra. de Valldemossa km. 7.5, E-07122 Palma, Spain
{jamila.mifdal,alessandro.sebastianelli}@esa.int, {joan.duran,tomeu.coll}@uib.es
marc.tomas1@estudiant.uib.es

Abstract

The fusion of multi-source data with different spatial and spectral resolutions is a crucial task in many remote sensing and computer vision applications. Model-based fusion methods are more interpretable and flexible than pure data-driven networks, but their performance depends greatly on the established fusion model and the hand-crafted prior. In this work, we propose an end-to-end trainable model-based network for hyperspectral and panchromatic image fusion. We introduce an energy functional that takes into account classical observation models and incorporates a high-frequency injection constraint. The resulting optimization function is solved by a forward-backward splitting algorithm and unfolded into a deep-learning framework that uses two modules trained in parallel to ensure both data observation fitting and constraint compliance. Extensive experiments are conducted on the remote-sensing hyperspectral PRISMA dataset and on the CAVE dataset, proving the superiority of the proposed deep unfolding network qualitatively and quantitatively.

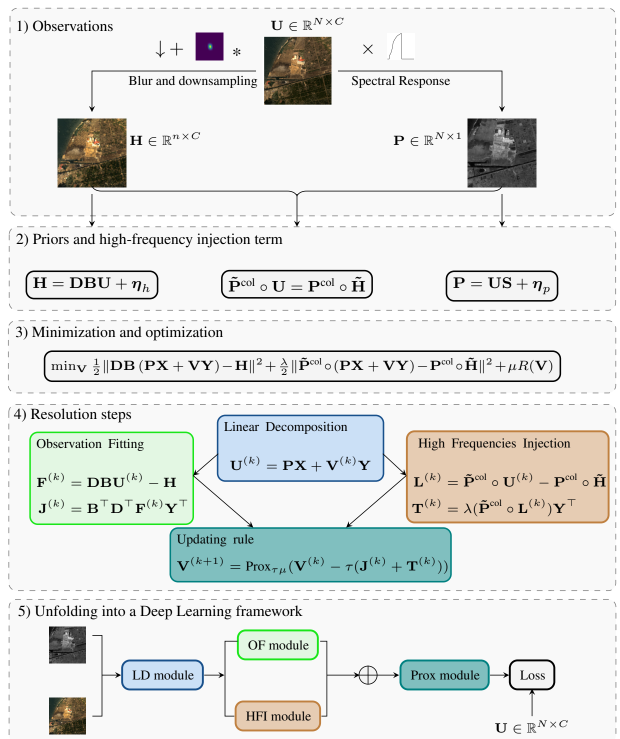

Figure 1. General scheme of the proposed hypersharpening method. After the generation of the observed images, the data fitting terms and the high-frequency injection constraint are extracted and construct the minimization function. The algorithmic steps are unfolded into a deep learning framework.

1. Introduction

Image fusion consists in gathering all relevant information from multiple images, which can be acquired by different devices, to usually produce a single one with better properties, such as high spatial and spectral resolutions. Due to the growing availability of satellite missions and cameras that capture multiple aspects of our everyday life, multi-source data fusion became an important technique for details recovery. A plethora of applications rely on the relevance of the image details for various downstream tasks such as Earth and environment monitoring [15,32,35], surveillance [11,21] or medical applications [20,26].

In the literature, pansharpening, which consists of the fusion of a multispectral image (MS) and a panchromatic one (PAN), has been widely investigated [10, 36, 37]. With the growing availability of hyperspectral (HS) sensors, hypersharpening, where HS data is used instead of the MS one, became a common method for hyperspectral spatial resolution enhancement [25, 27]. The fusion of HS and MS images has recently witnessed a growing interest [7, 34].

The growing popularity of deep learning (DL) has lead to an increase in fusion techniques [8, 18, 19, 28, 41]. However, in most cases, the architecture of DL methods is often not intuitive and is based solely on empirical justifications, which might cause problems such as vanishing gradient and not properly learning the scale difference between PAN and MS/HS images.

In this paper, we propose a deep unfolding model-driven method for hypersharpening, i.e., the fusion of a HS image and a PAN image. We focus on satellite imagery, but we show that our model is applicable to other types of computer vision datasets. The main contributions of this work can be summarized as follows:

- We introduce a new model-based method for hypersharpening using a high-frequency details injection constraint that extracts geometry and fine details from the PAN data and injects them in the fused image.
- The model is formulated as the minimization of an energy functional that is optimized using a forward-backward splitting algorithm. The solution is unfolded into a DL framework with a loss function that accounts for the high-frequency details injection constraint.
- The performance of our deep unfolding network is tested on images captured by the recent PRISMA mission[1] and on the CAVE database [42], and compared to that of traditional, recent model-based and end-to-end learned fusion algorithms.

The rest of the paper is organized as follows. In section 2, we review the state of the art (SOTA) on the fusion of data with different spatial and spectral resolutions. Section 3 introduces the proposed deep unfolding network. Its performance is exhaustively evaluated in Section 4 on PRISMA and CAVE datasets. An ablation study showing the potential of the high-frequency details injection module is also included. Conclusions are drawn in Section 5.

2. Related work

2.1. Classical methods

Hypersharpening and pansharpening methods are a subbranch of HS/MS image fusion. Both of them use the PAN image and increase the spatial resolution of either the MS image for pansharpening or the HS one for hypersharpening while perserving their spectral content. Thus in hypersharpening the number of spectral bands is much higher that in pansharpening. Most of the methods used for pansharpening could be easily adapted to hypersharpening. A way to classify traditional pansharpening methods is to group them in three main categories: variational methods [4, 9, 12], component-substitution (CS) algorithms [2, 13, 14] and multi-resolution analysis (MRA) techniques [1, 3].

[1] https://www.asi.it/en/earth-science/prisma

The CS methods are based on the fact that a spectral transformation is applied to the MS or HS image, and then, the spatial component is substituted with the PAN image. CS algorithms encompass Intensity Hue Saturation [14], Principal Component Analysis [13] and Gram-schmidt [2]. Regarding the MRA techniques, they rely on the extraction of spatial details, throughout a decomposition of the PAN image, which are injected in the MS bands [1, 3]. CS and MRA methods can suffer from spectral and spatial distortion during the details injection process which is linked to the choice of transformation and the type of decomposition of the PAN image.

As to the variational methods, they make the assumption that the PAN and the MS or the HS images are, respectively, a spectral and spatial degradation of the unknown image. From this, the fusion problem is formulated as the minimization of an energy function using some prior knowlegde [4, 9, 12]. The main drawback of the variational-based methods is the computational complexity due to the optimization process. Regarding the CS and the MRA methods, they can suffer from either spectral or spatial distortions because the extracted details depend on the chosen transformation and decomposition.

2.2. Deep learning based methods

In the last decade, a growing number of DL based pansharpening and hypersharpening methods, with various CNN architectures, were suggested in the literature and showed promising performances [8, 18, 19, 28, 41]. The MS or HS image and the PAN one are fed to the neural network and go through a succession of convolutional layers where the features and fine details are extracted, during this process the weights of the network are updated in order to fit the desired output.

The pioneer in this field was PNN [28] with a CNN-based architecture for the pansharpening task. PNN was built on the CNN-based model for super-resolution [8] and it is composed of three convolutional layers which makes it a basic neural network. In [19] the authors suggested the DiCNN network for learning the details from the PAN image and injects them in the pre-interpolated MS one in an end-to-end manner. Another network called PanNet [41], built upon the PNN architecture [28], used a ResNet [18] structure in order to improve the performance of the CNN model. Most of the DL based methods upsample the low-resolution input image to the size of the PAN one during the high-frequency feature extraction phase which introduces spectral distortions and does not make proper use of the information present in the low-resolution image.

2.3. Unfolding and model-driven based models

Most of the DL based methods for the fusion task in general are intuitively constructed with no justification behind the use of the network's structure. Also, when a CNN-based model does not provide the desired output, deepening it does not necessarily improve the results and most of the time leads to issues such as higher computational complexity, gradient disappearance, etc.

New model-based networks which are based on algorithm unfolding [31], made their entrance in the literature and proved very efficient in terms of performance. The idea of model-based methods is the formulation of an optimization function constructed with data observation models and priors about the desired output. The steps of the optimization algorithm are unfolded into a DL framework.

MHF-net [39] suggested a model-based HS/MS fusion network adaptable to hypersharpening tasks. The authors harnessed the low-rank property of the HS image in order to reduce the spectral distortion and unfolded the algorithm into convolutional layers. Another model-based network GPPNN [40] has two optimization problems, one for the PAN part and another one for the MS part. The two problems were solved separately and the unfolded networks ate stacked alternatively for the pansharpening task. The common factor between these model-based pansharpening and fusion mehtods is that the formulated optimization problem contains only the data-fitting terms that are extracted from the observation model and a regularizer to account for the ill-posedness of the optimization problem. Hence, the constructed network does not have the possibility to extract non-linear complex features from the data at hand.

3. Proposed unfolded hypersharpening method

Let $\mathbf{U} \in \mathbb{R}^{N\times C}$ be the target high-resolution HS image with $N = N_x \cdot N_y \in \mathbb{Z}^{>0}$ pixels and $C \in \mathbb{Z}^{>0}$ spectral bands, $\mathbf{H} \in \mathbb{R}^{n\times C}$ the low-resolution HS image with $n = \frac{N_x}{l}\cdot\frac{N_y}{l} \in \mathbb{Z}^{>0}$ pixels, where $l \in \mathbb{Z}^{>0}$ is the sampling factor, and $\mathbf{P} \in \mathbb{R}^{N\times 1}$ is the high-resolution PAN image. In this setting, it is assumed that $\mathbf{P}$ contains the high frequencies, i.e. the geometry of the scene being observed.

3.1. Hypersharpening model

The observation models [29, 30] relating $\mathbf{H}$ and $\mathbf{P}$ with $\mathbf{U}$ are generally given by

$$\begin{aligned} \mathbf{H} &= \mathbf{DBU} + \boldsymbol{\eta}_h, \\ \mathbf{P} &= \mathbf{US} + \boldsymbol{\eta}_p, \end{aligned} \tag{1}$$

where $\mathbf{B} \in \mathbb{R}^{N\times N}$ is the low-pass filter modeling the point spread function of the HS sensors, $\mathbf{D} \in \mathbb{R}^{n\times N}$ is the l-fold downsampling operator, $\mathbf{S} \in \mathbb{R}^{C\times 1}$ is the spectral response of the PAN sensor, and $\boldsymbol{\eta}_h$ and $\boldsymbol{\eta}_p$ are assumed to be additive, white Gaussian noise.

Usually, the linear operators $\mathbf{B}$, $\mathbf{D}$ and $\mathbf{S}$ can be obtained by registration and radiometric calibration. But, even when they are known, inferring $\mathbf{U}$ from (1) is an ill-posed inverse problem and additional priors are thus required. One can tackle the ill-posedness in the variational framework by introducing a regularization term $R(\mathbf{U})$ promoting smoothness of the solution. Then, $\mathbf{U}$ can be estimated by solving the following minimization problem:

$$\min_{\mathbf{U}} \frac{1}{2}\|\mathbf{DBU} - \mathbf{H}\|^2 + \frac{\gamma}{2}\|\mathbf{US} - \mathbf{P}\|^2 + \mu R(\mathbf{U}), \tag{2}$$

where $\|\cdot\|$ stands for the classical Frobenius norm and $\gamma, \mu > 0$ are trade-off parameters balancing the contribution of each term to the full energy.

The difference in the spatial resolution between the HS observation $\mathbf{H}$ and the PAN data $\mathbf{P}$ has to be captured accurately. A common approach consists in upscaling $\mathbf{H}$ to match the target resolution, but this might cause spectral distortions due to aliasing and impact the reconstruction of the fused image, specially when the sampling factor is relatively large. In order to avoid such issues while recovering the geometry of the scene, we introduce a constraint that injects the high-frequency details of $\mathbf{P}$ to the fused result.

On the one hand, the high frequencies of the fused image can be estimated as $\mathbf{U} - \tilde{\mathbf{H}}$, where $\tilde{\mathbf{H}} \in \mathbb{R}^{N\times C}$ is the result of upscaling $\mathbf{H}$ by bicubic interpolation. On the other hand, the high frequencies of the scene are given by $\mathbf{P} - \tilde{\mathbf{P}}$, where $\tilde{\mathbf{P}} \in \mathbb{R}^{N\times 1}$ contains the low frequencies of the PAN data. To get $\tilde{\mathbf{P}}$, we first apply the spatial degradation described in (1) to $\mathbf{P}$ and obtain a low-resolution image which is then upsampled by bicubic interpolation. We finally impose the high-frequency details injection constraint:

$$U_{ij} - \tilde{H}_{ij} = \frac{\tilde{H}_{ij}}{\tilde{P}_i}(P_i - \tilde{P}_i), \tag{3}$$

where $\frac{\tilde{H}_{ij}}{\tilde{P}_i}$ is a modelation coefficient that takes into account the energy levels of each spectral band. It is straightforward to see that (3) can be rewritten as

$$\tilde{\mathbf{P}}^{\text{col}} \circ \mathbf{U} = \mathbf{P}^{\text{col}} \circ \tilde{\mathbf{H}}, \tag{4}$$

where $\circ$ denotes the Hadamard (entrywise matrix) product and $\mathbf{P}^{\text{col}}, \tilde{\mathbf{P}}^{\text{col}} \in \mathbb{R}^{N\times C}$ are the replication of $\mathbf{P}$ and $\tilde{\mathbf{P}}$ to C columns, i.e., $P_{ij}^{\text{col}} = P_{ik}^{\text{col}}$ for all $j, k \in \{1, \dots, C\}$.

Before adding (4) to the energy, we also exploit the low-rankness prior structure along the spectral mode of the high-resolution HS image [33,39,44]. Accordingly, let us assume that $\mathbf{U}$ can be linearly represented by $\mathbf{P}$ and an unknown matrix $\mathbf{V} \in \mathbb{R}^{N\times(r-1)}$, where $r = \text{rank}(\mathbf{U}) > 1$, i.e.,

$$\mathbf{U} = \mathbf{PX} + \mathbf{VY} \tag{5}$$

with coefficient matrices $\mathbf{X} \in \mathbb{R}^{1\times C}$ and $\mathbf{Y} \in \mathbb{R}^{(r-1)\times C}$ to be learned. Therefore, the observation models (1) can be replaced by

$$\mathbf{H} = \mathbf{DB}\left(\mathbf{PX} + \mathbf{VY}\right) + \boldsymbol{\eta}, \tag{6}$$

where $\boldsymbol{\eta}$ denotes the noise, and the high-frequency details injection constraint (4) becomes

$$\tilde{\mathbf{P}}^{\text{col}} \circ (\mathbf{PX} + \mathbf{VY}) = \mathbf{P}^{\text{col}} \circ \tilde{\mathbf{H}}. \tag{7}$$

Putting it all together, the proposed fusion model is

$$\begin{aligned}\min_{\mathbf{V}}\ &\mu R(\mathbf{V}) + \frac{1}{2}\|\mathbf{DB}\,(\mathbf{PX} + \mathbf{VY}) - \mathbf{H}\|^2 \\ &+ \frac{\lambda}{2}\|\tilde{\mathbf{P}}^{\text{col}} \circ (\mathbf{PX} + \mathbf{VY}) - \mathbf{P}^{\text{col}} \circ \tilde{\mathbf{H}}\|^2,\end{aligned} \tag{8}$$

where $\lambda, \mu > 0$ are trade-off parameters and R is an arbitrary regularization term that will be learned. Note that we have applied the regularization on $\mathbf{V}$ instead of $\mathbf{U}$ to preserve the geometry of $\mathbf{P}$ in (5).

3.2. Forward-backward splitting algorithm

To solve (8), first note that the function

$$\begin{aligned}F(\mathbf{V}) = &\frac{1}{2}\|\mathbf{DB}\,(\mathbf{PX} + \mathbf{VY}) - \mathbf{H}\|^2 \\ &+ \frac{\lambda}{2}\|\tilde{\mathbf{P}}^{\text{col}} \circ (\mathbf{PX} + \mathbf{VY}) - \mathbf{P}^{\text{col}} \circ \tilde{\mathbf{H}}\|^2\end{aligned} \tag{9}$$

is differentiable, thus we may use the forward-backward splitting method [5, 6] to compute the solution. The basic idea is to combine an explicit step of descent in the smooth function F with an implicit step of descent in μR:

$$\mathbf{V}^{k+1} = \text{prox}_{\tau\mu}\left(\mathbf{V}^k - \tau\nabla F(\mathbf{V}^k)\right), \tag{10}$$

where $\tau > 0$ is the stepsize parameter, k the iteration number, and $\text{prox}_{\tau\mu}$ the proximity operator of μR, i.e.,

$$\text{prox}_{\tau\mu}(\hat{\mathbf{V}}) = \arg\min_{\mathbf{V}} \frac{1}{2\tau}\|\mathbf{V} - \hat{\mathbf{V}}\|^2 + \mu R(\mathbf{V}). \tag{11}$$

Since (11) behaves as a denoising energy of $\hat{\mathbf{V}}$, it can be replaced by a denoising network.

The final updating rule (10) is obtained by computing the differential of the smooth function F:

$$\begin{aligned}\nabla F(\mathbf{V}) = &\mathbf{B}^\top\mathbf{D}^\top\left(\mathbf{DB}\,(\mathbf{PX} + \mathbf{VY}) - \mathbf{H}\right)\mathbf{Y}^\top \\ &+ \lambda\,[\tilde{\mathbf{P}}^{\text{col}} \circ (\tilde{\mathbf{P}}^{\text{col}} \circ (\mathbf{PX} + \mathbf{VY}) - \mathbf{P}^{\text{col}} \circ \tilde{\mathbf{H}})]\mathbf{Y}^\top.\end{aligned} \tag{12}$$

The steps of the updating rule (12) are now unfolded into a DL framework.

3.3. Algorithm unfolding

The steps of the optimization algorithm (12) that solve the fusion problem (8) can be decomposed into four steps as highlighted in the left side of Figure 2. In the first step, the unknown image is represented with a linear decomposition involving the PAN data. Afterwards, two parallel steps consist of computing the terms related to the observation model and the high-frequency details injection, respectively. Finally, the last step consists of applying the proximal operator in order to update the variable of the minimization problem. Each one of these four steps are converted into a DL framework.

In this framework, we use the tensor formulations for all images to keep their spatial structure. Furthermore, we introduce the following operators:

- The operator $\texttt{Conv}_{b_{in}\longrightarrow b_{out}}$ takes a tensor with b_{in} bands and outputs a result with b_{out} bands.
- The operator $\texttt{dSamp}_{n_{in}\longrightarrow n_{out}}$ downsamples an input spatially from n_{in} to n_{out} pixels. It is composed of a blurring convolutional operator followed by a downsampling operation by a scale factor of $\frac{n_{in}}{n_{out}}$.
- The operator $\texttt{uSamp}_{n_{in}\longrightarrow n_{out}}$ spatially upsamples an input from n_{in} to n_{out} pixels.
- The operator $\texttt{pMult}(\mathcal{A}, \mathcal{B})$ carries out a point-wise multiplication between the tensors $\mathcal{A}$ and $\mathcal{B}$.
- The operator $\texttt{ProxNet}$ stands for the proximal operator and it is replaced by a ResNet [18] as suggested in [39].

The proximity operator of R can be equivalently defined as a resolvent operator [5], i.e., $\text{prox}_{\tau\mu} = (\text{Id} + \tau\mu\partial R)^{-1}$, therefore, it is typically close to the identity. This is one of the reasons why we use a residual network to encode the proximity operator. The second main reason is that these networks are easier to train because they only need to learn a small offset from the identity.

The right side of Figure 2 shows the corresponding operations in the DL framework of the hypersharpening algorithm. The four blocs highlighted in Figure 2 are the main components of the complete network illustrated in Figure 3 where we could see three main stages. Each stage is composed of the blocks detailed in Figure 2 and at each epoch all three stages are executed, then, the estimated result is fed to the loss function.

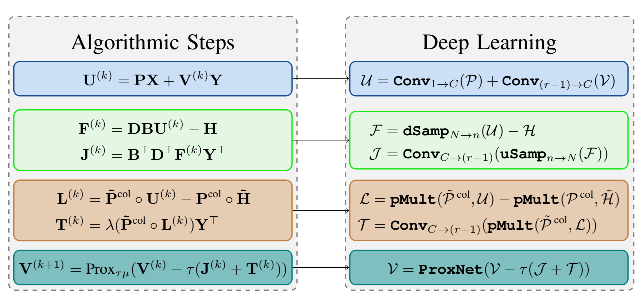

Figure 2. Relationship between the steps of the optimization algorithm and the modules of the deep unfolded network.

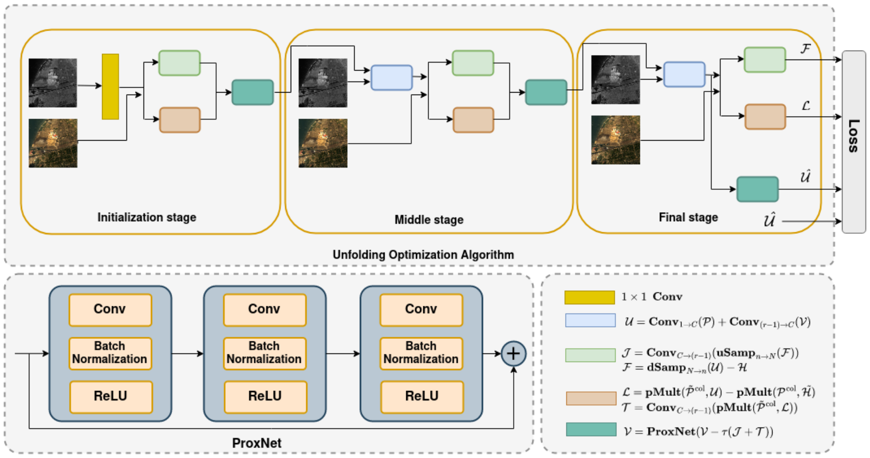

Figure 3. The unfolded network of the hypersharpening algorihtm using a high-frequency injection module. The network is composed of three stages, each one of these stages follows the unfolding steps detailed in Figure 2.

3.4. Training details

The proposed deep unfolding network is trained using the following loss function:

$$L = \|\hat{\mathbf{U}}^{(k)} - \mathbf{U}\|^2 + \alpha\|\mathbf{F}^{(k)}\|^2 + \beta\|\mathbf{L}^{(k)}\|^2, \qquad (13)$$

where $\hat{\mathbf{U}}^{(k)}$ is the estimated fused image at each epoch, $\mathbf{F}^{(k)}$ and $\mathbf{L}^{(k)}$ are respectively taken from observation-fitting and high-frequency details injections steps, and α and β are trade-off parameters. The first term of the loss function is an L^2 norm between the solution proposed by the network and the reference. The second term accounts for the residual error from the observation model and the last term represents the error when violating the high-frequency details injection module.

Our model is trained in a PyTorch framework, using an Nvidia A100 GPU, during 4000 epochs for the PRISMA dataset and 3000 for the CAVE dataset. We use an Adam optimizer with a learning rate of 10^{-3} and a batch size of 8 images. The trade-off parameters α and β were optimized and set to 10^{-3}. The images of both PRISMA and CAVE datasets were normalised by dividing on $2^{16} - 1$ and no augmentation techniques were applied.

4. Experiments

We conducted multiple experiments and compared the performances of our algorithm with the pure DL methods MSDCNN [45] and DiCNN [19], the deep unfolding networks MHFnet [39] and GPPNN [40], and classical fusion methods such as PCA [22], Brovey [16], GS [23], GSA [2], IHS [17] and SFIM [24].

For an objective comparison with the SOTA, we used the following qualitative metrics: PSNR (Peak Signal to Noise Ratio), which measures the reconstruction of the image quality with respect to noise, ERGAS (Erreur Relative Globale Adimentionnelle de Synthèse) and SSIM (Structural Similarity Index Measure), which measure the general quality of the fused image, and DD (Distortion Degree) and SAM (Spectral Angle Mapper), which measure the spectral reconstruction quality of the output image. We refer the reader to [43] for more details about the above indices.

Our model was tested on the recent PRISMA dataset and on the CAVE database [42]. For the simulation of the observation images we followed the Wald protocole [38]. For the experiments on PRISMA, the spatial downsampling operator $\mathbf{B}$ and the spectral degradation operator $\mathbf{S}$ were provided by the PRISMA mission engineers. Regarding the experiments on CAVE, the spatial and spectral operators were taken from available resources in the research community.

4.1. Experiments on PRISMA dataset

The PRISMA mission[1] was launched in 2019 by the Italian Space Agency (ASI), and to the best of our knowledge, it is the first mission that provides public HS and PAN data of the same region and presents substantial potential for fusion and resolution enhancement. The HS data has a spatial resolution of 30 m and contains 240 bands that cover the VNIR (Visible and Near Infra-Red) range: 400-1010 nm and the SWIR (Short Wave- length Infra-Red)) one: 20–2505 nm. The PAN data contains one single band at a spatial resolution of 5 m. We selected and downloaded 20 large-scale scenes of PRISMA images throughout the PRISMA mission's portal[2]. The downloaded HS images have an original size of $1000 \times 1000 \times 240$, each one of the scenes was cropped into non-overlapping tiles of $128 \times 128 \times 240$. Given that the SWIR bands are not covered by the spectral response of the PAN sensor, only the first 66 bands were considered which resulted in tiles

[2]Prisma portal: https://prisma.asi.it

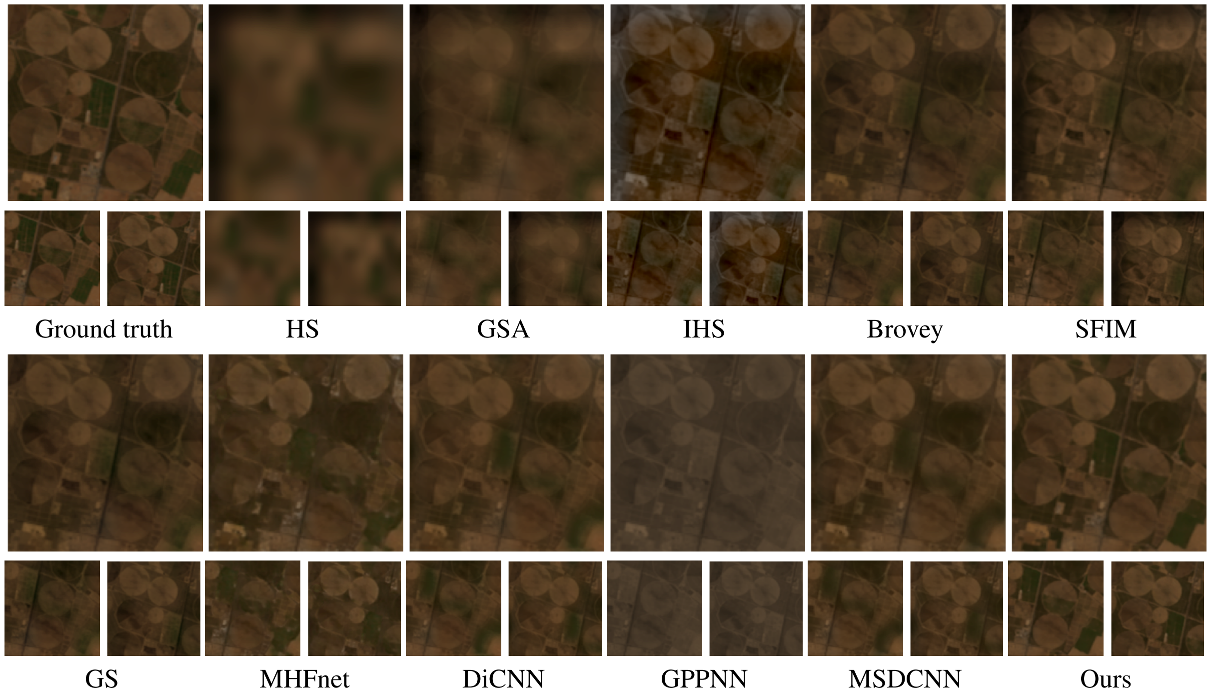

Figure 4. Visual comparison of the fusion approaches on an image of the PRISMA dataset. We display the 35th, 45th and the 57th bands in place of the RGB channels. The proposed deep unfolding network successfully combines the geometry of the PAN image with the spectral information of the HS data, while all other results are affected by blur, color artifacts and spatial distortions. Our method is also able to recover both large structures, such as the circular ground contours, and the finest ones, such as roads and small building structures.

of $128 \times 128 \times 66$, from each tile a new HS and PAN images were generated following the Wald protocol [38] and using the spectral and spatial responses provided by PRISMA mission engineers. The chosen downsampling factor for the PRISMA dataset is 12, thus, from each tile of $128 \times 128 \times 66$, an HS image of $11 \times 11 \times 66$ and PAN image of the size 128×128 were considered for the hypersharpening process.

For the training process 640 tiles from PRISMA scenes were used and 128 tiles were utilized for the validation step. The training and the validation dataset were from different regions in order to test the model's ability to generalize to unseen regions.

Table 1 displays the average of the quality measures obtained for each fusion method over all images of the PRISMA dataset. The best results are in bold and the second best ones are underlined. We observe that the proposed deep unfolding network significantly outperforms all the others with respect to all metrics. Interestingly, the pure DL approaches DiCNN and MSDCNN give better quantitative results than the unfolding networks GPPNN and MHFnet, while our method clearly outperforms all of them. This proves the suitability of the proposed hypersharpening model (8) among deep unfolding strategies, providing a significant increase in terms of spectral and spatial qualities.

Table 1. Average of the quality measures over all images of the PRISMA dataset. The methods are divided into classical, pure DL and deep unfolding categories. The best results are in bold and the second best ones are underlined. We observe that the proposed deep unfolding network significantly outperforms all the other fusion methods with respect to all quantitative metrics.

	ERGAS ↓	PSNR ↑	SSIM ↑	DD ↓	SAM ↓
PCA	376.06	15.70	0.3990	0.1333	35.41
Brovey	92.86	27.68	0.9180	0.0309	4.69
Bicubic	225.81	23.40	0.8303	0.0420	4.64
GS	92.14	27.75	0.9181	0.0308	4.78
GSA	186.01	23.98	0.8706	0.0399	4.67
IHS	101.00	26.78	0.8882	0.0346	7.20
SFIM	208.91	23.95	0.8801	0.0392	4.67
DiCNN	41.44	33.38	0.9520	0.0145	3.70
MSDCNN	43.10	33.07	0.9496	0.0153	4.06
GPPNN	253.18	20.99	0.8453	0.0700	7.92
MHFnet	45.29	32.69	0.9402	0.0157	4.13
Ours	**15.31**	**42.17**	**0.9900**	**0.0078**	**1.59**

Figure 4 displays the fused images obtained by each technique putting the 35th, 45th and the 57th bands in place of the RGB channels. All SOTA methods are affected by

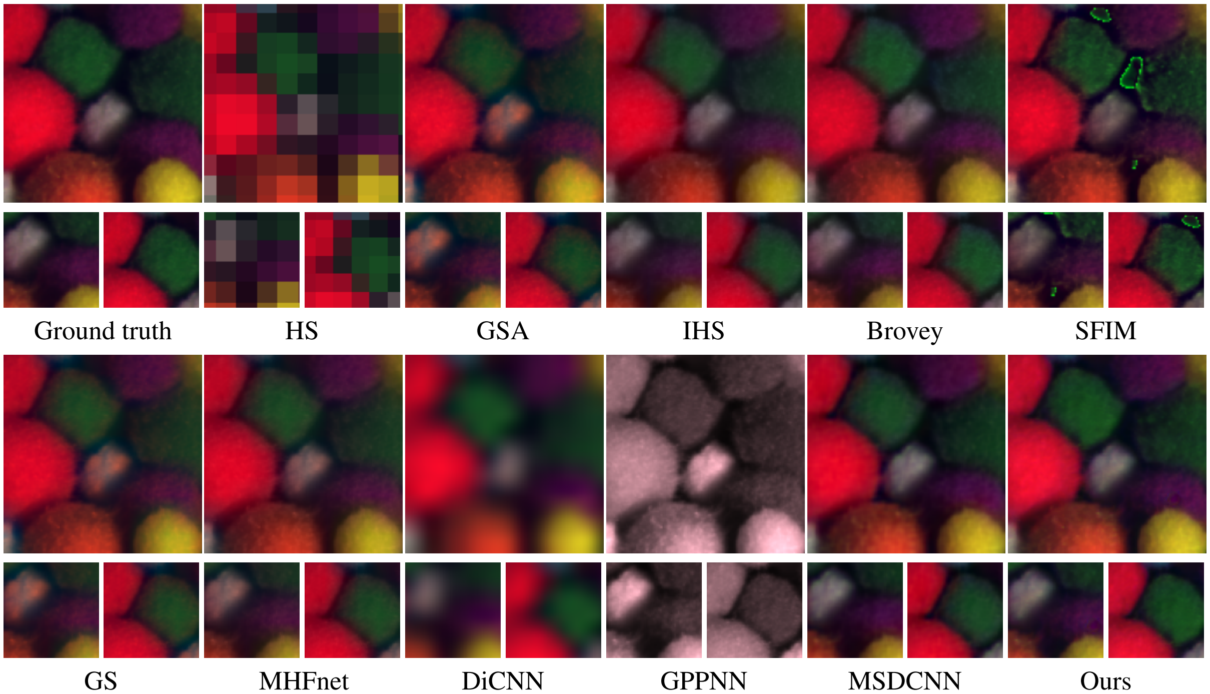

Figure 5. Visual comparison of the fusion approaches on CAVE dataset, we use the 28th, 13th and the first bands in place of the RGB ones. DL methods like DiCNN, MSDCNN and MHFnet suffer from blurring and others suffer from spectral artifacts visible on the left highlighted parts, also GPPNN did not manage to recover the spectral information. Our deep unfolding network gives the best visual result in terms of spectral consistency and geometry retrieval.

blur, color artifacts and spatial distortions, giving rise to fused images which are not visually pleasant. On the contrary, our deep unfolding approach is able to correctly preserve the spectral information from the HS data while injecting the geometry from the PAN image. Furthermore, we observe that the proposed method is able to recover large structures, such as the circular ground contours, as well as the finest ones, such as roads and small building structures.

4.2. Experiments on CAVE dataset

We test our network on the CAVE dataset [42] which is composed of 32 scenes with the original size $512 \times 512 \times 31$. From each scene crops of $128 \times 128 \times 31$ were extracted and used to generate a new HS images of size $11 \times 11 \times 31$ and PAN images of size 128×128 following the Wald protocol [38] and using a downsampling factor of 12.

Table 2 displays the average of the quality measures over all images of the CAVE dataset. The best results are in bold and the second best ones are underlined. The proposed fusion method significantly outperforms the other techniques in terms of all the metrics. Figure 5 shows the hypersharpening results of our networks and of the SOTA methods. We notice that DL methods like DiCNN, MSDCNN and MHFnet suffer from blurring, other techniques suffer from spectral artifacts visible on the left highlighted parts and GPPNN did not manage to recover the spectral information. Our deep unfolding network gives the best visual result in terms of spectral consistency and geometry retrieval which shows the importance of the high-frequency injection mod-

Table 2. Average of the quality measures over all images of the CAVE dataset. The methods are divided into classical, pure DL and deep unfolding categories. The best results are in bold and the second best ones are underlined. Our deep unfolding network outperforms all the others with respect to all quantitative metrics.

	ERGAS ↓	PSNR ↑	SSIM ↑	DD ↓	SAM ↓
PCA	262.32	18.4060	0.6759	0.0951	22.2876
Brovey	67.12	30.0125	0.9533	0.0231	5.3162
Bicubic	92.81	27.1778	0.9009	0.0278	4.5678
GS	83.97	28.0502	0.9208	0.0285	6.8994
GSA	73.38	29.3597	0.9296	0.0230	6.1939
IHS	78.53	28.7951	0.9340	0.0272	6.4508
SFIM	110.20	25.8364	0.9134	0.0271	6.0997
DiCNN	92.70	27.1756	0.9009	0.0278	4.5602
MSDCNN	46.50	33.9718	0.9655	0.0144	4.3489
GPPNN	508.38	15.7420	0.6478	0.1068	24.7192
MHFnet	81.25	28.4226	0.9274	0.0278	6.6751
Ours	**40.06**	**34.7503**	**0.9695**	**0.0134**	**4.3110**

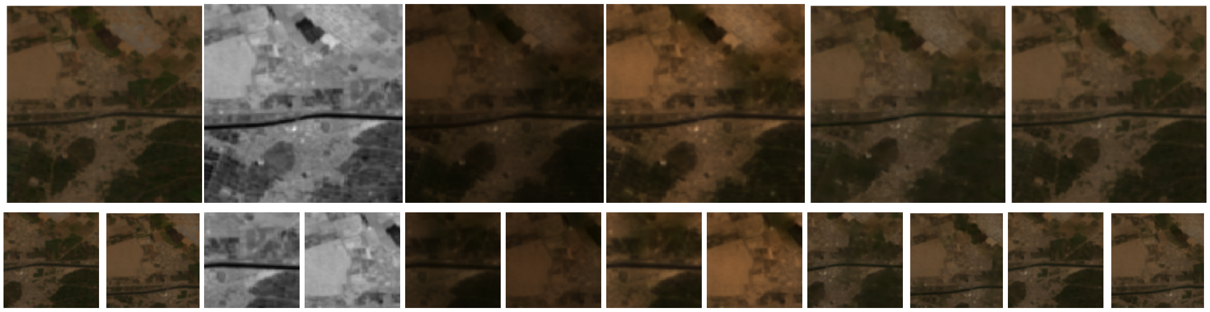

Figure 6. Visual comparison on a PRISMA crop for the ablation study. The SOTA methods and our network without the high-frequency (HF) injection module have a poor reconstruction of the colors and, for instance, fail in detecting the white road appearing in the left-hand crop. On the contrary, the full proposed deep unfolding network is able to reconstruct the spatial and spectral information of the scene.

ule in copying the accurate information from the HS and PAN images.

4.3. Ablation study

In this part we tested the importance of the introduced high-frequency (HF) details injection module. For this purpose, we trained the model only with the observation-fitting term on the same training and validation dataset used for the suggested fusion method on PRISMA dataset. The reason behind using PRISMA dataset is because it has a much lower resolution than CAVE thus it is more challenging to recover high-frequency details such as the contours and the fine geometrical information.

Figure 6 shows the result on a PRISMA image that contains both high-frequency details such as roads and uniform structure such as green fields. The results were compared to the best SOTA methods in terms of objective performances. We can see that, on the highlighted parts on the left, the result produced by the model without high-frequency module “No HF module”) and the fused images from the SOTA methods, failed in detecting the little white road, that crosses the bushes, except ours that reconstructed it accurately. Also, all the fused images had a poor reconstruction of the colors in multiple parts of the image whereas our result recovered the colors with minimal artifacts.

The visual observation are confirmed by the objective results showed in Table 3, where our method outperforms all the others. We can conclude that the high-frequency injection module has a crucial role in recovering the texture and the fine geometrical details of the fused images.

Table 3. Quality measures on a PRISMA crop for the ablation study. The best results are in bold and the second best ones are underlined. The proposed deep unfolding network outperforms all the others. The indices also prove the relevance of the high-frequency details injection module in the fusion process.

	ERGAS ↓	PSNR ↑	SSIM ↑	DD ↓	SAM ↓
DiCNN	<u>41.44</u>	<u>34.70</u>	<u>0.9562</u>	<u>0.0126</u>	<u>4.04</u>
MSDCNN	43.36	34.46	0.9546	0.0132	4.39
No HF	46.51	33.78	0.9481	0.0141	4.59
Ours	**20.89**	**41.27**	**0.9905**	**0.0063**	**2.39**

5. Conclusions

In this paper we proposed a novel model-based neural network for hypersharpening. The model takes advantage of the observation data and uses a high-frequency details injection term. The algorithmic steps obtained from the resolution of a minimization problem are unrolled into a DL framework. The experiments were conducted on two types of datasets. On the recent remote-sensing PRISMA dataset the hypersharpening model proved its ability in recovering the fine details. We also tested the performance of the suggested network on the CAVE dataset which has a higher spatial resolution and the results were competitive with respect to the SOTA methods, which shows the generalizability of the model to different resolutions. We also emphasized the importance of the introduced high-frequency details injection module in reconstructing the fine spatial and spectral details in an ablation study.

Acknowledgements

This work is part of the MaLiSat project TED2021-132644B-I00 funded by MCIN/AEI/10.13039/501100011033 and by the European Union “NextGenerationEU”/PRTR. The authors were also supported by the Conselleria de Fons Europeus, Universitat i Cultura of the Govern de les Illes Balears under grant AP_2021_023. The authors would like to thank warmly the PRISMA mission engineers for providing the sensors responses.

References

[1] Bruno Aiazzi, L Alparone, Stefano Baronti, Andrea Garzelli, and Massimo Selva. Mtf-tailored multiscale fusion of high-resolution ms and pan imagery. *Photogrammetric Engineering & Remote Sensing*, 72(5):591–596, 2006. 2

[2] Bruno Aiazzi, Stefano Baronti, and Massimo Selva. Improving component substitution pansharpening through multivariate regression of ms + pan data. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10):3230–3239, 2007. 2, 5

[3] Arian Azarang and Hassan Ghassemian. A new pansharpening method using multi resolution analysis framework and deep neural networks. In *2017 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA)*, pages 1–6. IEEE, 2017. 2

[4] Coloma Ballester, Vicent Caselles, Laura Igual, Joan Verdera, and Bernard Rougé. A variational model for p+ xs image fusion. *International Journal of Computer Vision*, 69(1):43–58, 2006. 2

[5] Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016. 4

[6] Patrick L Combettes and Valérie R Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale modeling & simulation*, 4(4):1168–1200, 2005. 4

[7] Renwei Dian, Shutao Li, Bin Sun, and Anjing Guo. Recent advances and new guidelines on hyperspectral and multispectral image fusion. *Information Fusion*, 69:40–51, 2021. 1

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 1, 2

[9] Joan Duran, Antoni Buades, Bartomeu Coll, and Catalina Sbert. A nonlocal variational model for pansharpening image fusion. *SIAM Journal on Imaging Sciences*, 7(2):761–796, 2014. 2

[10] Joan Duran, Antoni Buades, Bartomeu Coll, Catalina Sbert, and Gwendoline Blanchet. A survey of pansharpening methods with a new band-decoupled variational model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 125:78–105, 2017. 1

[11] Sara Freitas, Hugo Silva, José Miguel Almeida, and Eduardo Silva. Convolutional neural network target detection in hyperspectral imaging for maritime surveillance. *International Journal of Advanced Robotic Systems*, 16(3):1729881419842991, 2019. 1

[12] Xueyang Fu, Zihuang Lin, Yue Huang, and Xinghao Ding. A variational pan-sharpening with local gradient constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10265–10274, 2019. 2

[13] Mohamed Ghadjati, Abdelkrim Moussaoui, and Abdelhak Boukharouba. A novel iterative pca–based pansharpening method. *Remote sensing letters*, 10(3):264–273, 2019. 2

[14] Morteza Ghahremani and Hassan Ghassemian. Nonlinear ihs: A promising method for pan-sharpening. *IEEE Geoscience and Remote Sensing Letters*, 13(11):1606–1610, 2016. 2

[15] Marco Gianinetto and Giovanmaria Lechi. The development of superspectral approaches for the improvement of land cover classification. *IEEE Transactions on Geoscience and Remote Sensing*, 42(11):2670–2679, 2004. 1

[16] Alan R Gillespie, Anne B Kahle, and Richard E Walker. Color enhancement of highly correlated images. ii. channel ratio and "chromaticity" transformation techniques. *Remote Sensing of Environment*, 22(3):343–365, 1987. 5

[17] R Haydn. Application of the ihs color transform to the processing of multisensor data and image enhancement. In *Proc. of the International Symposium on Remote Sensing of Arid and Semi-Arid Lands, Cairo, Egypt, 1982*, 1982. 5

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2, 4

[19] Lin He, Yizhou Rao, Jun Li, Jocelyn Chanussot, Antonio Plaza, Jiawei Zhu, and Bo Li. Pansharpening via detail injection based convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(4):1188–1204, 2019. 1, 2, 5

[20] Garima Jaiswal, Arun Sharma, and Sumit Kumar Yadav. Critical insights into modern hyperspectral image applications through deep learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(6):e1426, 2021. 1

[21] Alper Koz. Ground-based hyperspectral image surveillance systems for explosive detection: Part i—state of the art and challenges. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(12):4746–4753, 2019. 1

[22] P Kwarteng and A Chavez. Extracting spectral contrast in landsat thematic mapper image data using selective principal component analysis. *Photogramm. Eng. Remote Sens*, 55(1):339–348, 1989. 5

[23] Craig A Laben and Bernard V Brower. Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening, Jan. 4 2000. US Patent 6,011,875. 5

[24] JG Liu. Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details. *International Journal of Remote Sensing*, 21(18):3461–3472, 2000. 5

[25] Laetitia Loncan, Luis B De Almeida, José M Bioucas-Dias, Xavier Briottet, Jocelyn Chanussot, Nicolas Dobigeon, Sophie Fabre, Wenzhi Liao, Giorgio A Licciardi, Miguel Simoes, et al. Hyperspectral pansharpening: A review. *IEEE Geoscience and remote sensing magazine*, 3(3):27–46, 2015. 1

[26] Guolan Lu and Baowei Fei. Medical hyperspectral imaging: a review. *Journal of biomedical optics*, 19(1):010901, 2014. 1

[27] Xiaochen Lu, Junping Zhang, Xiangzhen Yu, Wenming Tang, Tong Li, and Ye Zhang. Hyper-sharpening based on spectral modulation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(5):1534–1548, 2019. 1

[28] Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva, and Giuseppe Scarpa. Pansharpening by convolutional neural networks. *Remote Sensing*, 8(7):594, 2016. 1, 2

[29] Jamila Mifdal, Bartomeu Coll, Jacques Froment, and Joan Duran. Variational fusion of hyperspectral data by non-local filtering. *Mathematics*, 9(11):1265, 2021. 3

[30] Rafael Molina, Aggelos K Katsaggelos, and Javier Mateos. Bayesian and regularization methods for hyperparameter estimation in image restoration. *IEEE Transactions on Image Processing*, 8(2):231–246, 1999. 3

[31] Vishal Monga, Yuelong Li, and Yonina C Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021. 3

[32] R Neville. Automatic endmember extraction from hyperspectral data for mineral exploration. In *International Airborne Remote Sensing Conference and Exhibition, 4 th/21 st Canadian Symposium on Remote Sensing, Ottawa, Canada*, 1999. 1

[33] Zahra Hashemi Nezhad, Azam Karami, Rob Heylen, and Paul Scheunders. Fusion of hyperspectral and multispectral images using spectral unmixing and sparse coding. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(6):2377–2389, 2016. 3

[34] Dioline Sara, Ajay Kumar Mandava, Arun Kumar, Shiny Duela, and Anitha Jude. Hyperspectral and multispectral image fusion techniques for high resolution applications: A review. *Earth Science Informatics*, 14(4):1685–1705, 2021. 1

[35] Mary B Stuart, Andrew JS McGonigle, and Jon R Willmott. Hyperspectral imaging in environmental monitoring: a review of recent developments and technological advances in compact field deployable systems. *Sensors*, 19(14):3071, 2019. 1

[36] Gemine Vivone, Luciano Alparone, Jocelyn Chanussot, Mauro Dalla Mura, Andrea Garzelli, Giorgio A Licciardi, Rocco Restaino, and Lucien Wald. A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5):2565–2586, 2014. 1

[37] Gemine Vivone, Mauro Dalla Mura, Andrea Garzelli, Rocco Restaino, Giuseppe Scarpa, Magnus O Ulfarsson, Luciano Alparone, and Jocelyn Chanussot. A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods. *IEEE Geoscience and Remote Sensing Magazine*, 9(1):53–81, 2020. 1

[38] Lucien Wald, Thierry Ranchin, and Marc Mangolini. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogrammetric engineering and remote sensing*, 63(6):691–699, 1997. 5, 6, 7

[39] Qi Xie, Minghao Zhou, Qian Zhao, Deyu Meng, Wangmeng Zuo, and Zongben Xu. Multispectral and hyperspectral image fusion by ms/hs fusion net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1585–1594, 2019. 3, 4, 5

[40] Shuang Xu, Jiangshe Zhang, Zixiang Zhao, Kai Sun, Junmin Liu, and Chunxia Zhang. Deep gradient projection networks for pan-sharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1366–1375, 2021. 3, 5

[41] Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. Pannet: A deep network architecture for pan-sharpening. In *Proceedings of the IEEE international conference on computer vision*, pages 5449–5457, 2017. 1, 2

[42] Fumihito Yasuma, Tomoo Mitsunaga, Daisuke Iso, and Shree K Nayar. Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum. *IEEE transactions on image processing*, 19(9):2241–2253, 2010. 2, 5, 7

[43] Naoto Yokoya, Claas Grohnfeldt, and Jocelyn Chanussot. Hyperspectral and multispectral data fusion: A comparative review of the recent literature. *IEEE Geoscience and Remote Sensing Magazine*, 5(2):29–56, 2017. 5

[44] Naoto Yokoya, Takehisa Yairi, and Akira Iwasaki. Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 50(2):528–537, 2011. 3

[45] Qiangqiang Yuan, Yancong Wei, Xiangchao Meng, Huanfeng Shen, and Liangpei Zhang. 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, 2018. 5