Suplementary material: "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

In this supplementary material we provide more details about the hypersharpening algorithm, the implementation of the unfolded network along with information about the deep-learning settings and additional visual and objective results.

1. The fusion algorithm and unfolded network

The steps of the optimization algorithm that solves the fusion problem (8) of the main paper is detailed in Algorithm 1, these steps are then unfolded into a deep learning framework. The unfolded network shown in Figure 3 of the main paper is composed of three main stages and each stage follows the four steps described in Figure 2 of the main paper. In the initialization stage, the variable $\mathbf{V}$, with respect to which the minimization is carried out, is initialized with zeros. This implies that in the first iteration the unknown fused image $\mathbf{U}$ is computed as the 1×1 convolution of the image $\mathbf{P}$ which conserves the spatial size and increases the spectral one to C channels. In the middle stage, the output of the previous one along with the input data $\mathbf{P}$ are fed to the linear decomposition module. The result, along with the input image $\mathbf{H}$ are passed to both the observation-fitting and the high-frequency injection modules in parallel which produces the input to the last stage. In the latter, the output of the linear decomposition module goes through a **`ProxNet`** module in order to produce an estimate of the fused image which, along with the outputs of the data-fitting and the high-frequency injection modules, are fed to the loss function.

The module **`ProxNet`** is a residual network composed of three stages, each stage is a sequence of a convolution, batch normalization and a ReLu activation function as highlighted in Figure 3 of the main paper. The operator

Algorithm 1: Fusion optimization algorithm

Input: Observation data: $\mathbf{P}$ and $\mathbf{H}$, high-frequency injection terms: $\mathbf{P}^{\text{col}}$ and $\tilde{\mathbf{P}}^{\text{col}}$, hyper-parameters: $\lambda, \mu, \tau > 0$

for $k \leftarrow 0$ **to** *niters* **do**

- ***Linear decomposition***

$$\mathbf{U}^{(k)} = \mathbf{P}\mathbf{X} + \mathbf{V}^{(k)}\mathbf{Y}$$

- ***Observation fitting***

$$\mathbf{F}^{(k)} = \mathbf{D}\mathbf{B}\mathbf{U}^{(k)} - \mathbf{H}$$
$$\mathbf{J}^{(k)} = \mathbf{B}^{\top}\mathbf{D}^{\top}\mathbf{F}^{(k)}\mathbf{Y}^{\top}$$

- ***High frequencies injection***

$$\mathbf{L}^{(k)} = \tilde{\mathbf{P}}^{\text{col}} \circ \mathbf{U}^{(k)} - \mathbf{P}^{\text{col}} \circ \tilde{\mathbf{H}}$$
$$\mathbf{T}^{(k)} = \lambda(\tilde{\mathbf{P}}^{\text{col}} \circ \mathbf{L}^{(k)})\mathbf{Y}^{\top}$$

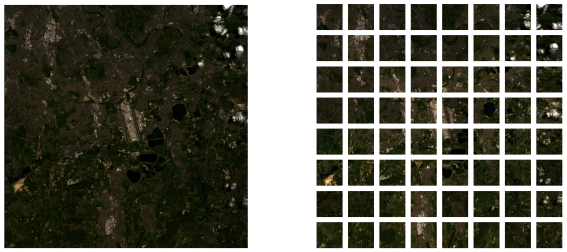
- ***Updating rule***

$$\mathbf{V}^{(k+1)} = \text{Prox}_{\tau\mu}\left(\mathbf{V}^{(k)} - \tau\left(\mathbf{J}^{(k)} + \mathbf{T}^{(k)}\right)\right)$$

end

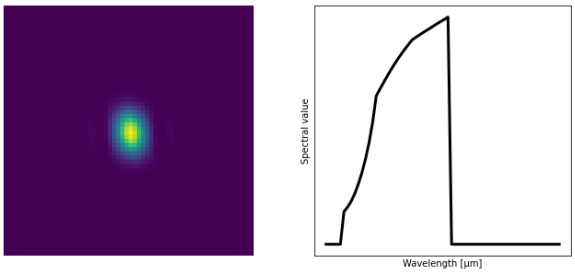
Output: $\mathbf{V}$

`dSamp`$_{n_{in} \longrightarrow n_{out}}$ downsamples an input spatially from n_{in} to n_{out} pixels. In the case of PRISMA dataset, the downsampling factor was chosen to be 12. In order to preserve as many details as possible, the downsampling was decomposed into three sub-downsamplings of factors two, three and two. For each sub-downsampling operation, a convolution was applied to respect the Shanon-Nyquist condition. Regarding the operator **`uSamp`**$_{n_{in} \longrightarrow n_{out}}$, it spatially upsamples an input from n_{in} to n_{out} pixels. This operator is delicate and could lead to a loss of fine details if not handled carefully. Thus, just like the downsampling operator, the upsampling opeartions took place in three sub-upsampling ones with factors of two, three and two. Each one of the



(a) (b)

Figure 1. Visualization of the data splitting for training and validation. (a) is the original HS 1000x1000x66 image and (b) shows the splitting into tiles of 128x128x66.



(a) (b)

Figure 2. The spatial and spectral responses used for HS and PAN data generation from the PRISMA mission. (a) is the spatial low-pass filter applied before downsampling and (b) is the spectral response of the PAN sensor.

sub-upsampling operations is represented with a transposed convolutional layer followed by a succession of, convolution, batch normalisation and ReLu in order to ensure a maximum preservation of details and a smooth transition to higher resolutions, a final convolution is applied to the output.

2. Implementation details

The implementation of the hypersharpening code was carried out using PyTorch, the parameters α and β were optimized using the PyTorch library Ray Tune and they were both set to 10^{-3} for the PRISMA and the CAVE datasets. The images of both datasets were normalised by dividing on $2^{16}-1$ because they are encoded on 16 bits and no augmentation techniques were applied.

The transposed convolutional filters in the hypersharpening code were initialized using a zero-mean Gaussian distribution with a standard deviation of 0.1. The operation `1x1 Conv` in the initialization stage (Figure 3 of the main paper) was initialized using:

$$\mathbf{I} = (\hat{\mathbf{P}}^2)^{-1}\hat{\mathbf{P}}^\top\tilde{\mathbf{U}}, \tag{1}$$

where $\mathbf{I}$ is the solution of the following least square problem:

Table 1. Average of the quality measures over the fours images from the PRISMA dataset displayed in Figure 3. 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	357.21	16.06	0.3014	0.1290	35.26
Brovey	81.57	28.91	0.9174	0.0270	4.86
Bicubic	235.36	22.93	0.7674	0.0460	4.90
GS	81.01	28.97	0.9176	0.0269	4.90
GSA	210.21	23.26	0.7733	0.0449	4.86
IHS	98.13	27.15	0.8845	0.0326	7.55
SFIM	193.68	24.21	0.8753	0.0379	4.81
DiCNN	<u>45.26</u>	<u>33.07</u>	<u>0.9364</u>	<u>0.0156</u>	<u>4.28</u>
MSDCNN	46.65	32.86	0.9347	0.0160	4.51
GPPNN	259.94	21.23	0.8351	0.0683	7.28
MHFnet	48.60	32.50	0.9290	0.0169	4.69
Ours	**19.48**	**40.39**	**0.9887**	**0.0070**	**2.07**

$$\min_{\mathbf{I}} \|\hat{\mathbf{P}}\mathbf{I} - \tilde{\mathbf{U}}\|, \tag{2}$$

and $\hat{\mathbf{P}}$ and $\tilde{\mathbf{U}}$ are stacks, along the spatial dimension of PAN and reference images from the training set.

3. More results on PRISMA

In this section we include more results with the PRISMA dataset in order to show the ability of our unfolded network in recovering relevant spatial and spectral details, even when the observed images have a low-spatial resolution due to the downsampling factor which is 12 in our case. We start from the original HS data with the size of $1000 \times 1000 \times 66$ which has a spatial resolution of 30 m, then, the image is split into non-overlapping tiles of $128 \times 128 \times 66$ as highlighted in Figure 1. From each tile a new HS and PAN images are generated following the Wald protocole [1] and using the spatial and spectral responses, shown in Figure 2, provided by the engineers from the PRISMA mission.

In Figure 1 we show the result of the proposed unfolded fusion network on various images from the PRISMA dataset using the 35th, 45th and the 57th bands in place of the RGB channels and we compare the performances of our result to the best three SOTA methods in terms of objective metrics. We can notice that, visually, our result looks similar to the ground truth and does not contain artifacts. On the contrary, all the SOTA methods, starting from the top row to the bottom, either contain artifacts like the MHFnet method in the second row, or they do not recover relevant geometrical

details like the green pond in the case of DiCNN and MS-DCNN in the last row. Whenever the SOTA methods failed to reconstruct a spatial or spectral information our method managed to recover it, which shows the important role of the high-frequency details injection module.

In Table 1 we display the average of the quality measures over the four images displayed in Figure 1. The best results are in bold and the second best ones are underlined. We notice that the result of the proposed unfolding network significantly outperforms all the SOTA methods, especially the deep-unfolding ones, in terms of all the quality measures. The second best method is DiCNN which is a pure DL approach. These measures show the superiority of our deep-unfolding method in recovering the spatial and spectral details from an objective perspective.

References

[1] 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. 2

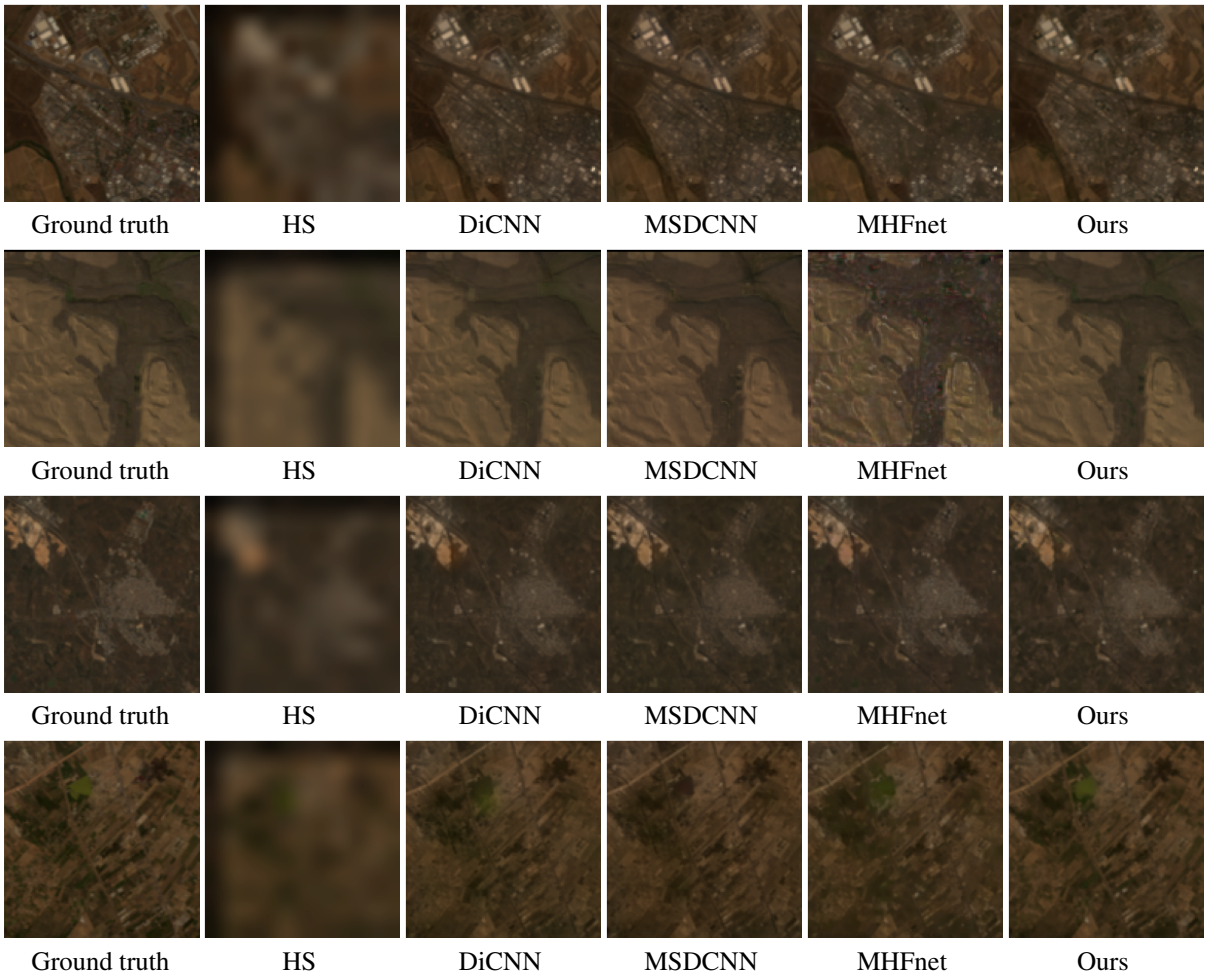


Figure 3. Visual comparison of the fusion approaches on four images from the PRISMA dataset. We display the 35th, 45th and the 57th bands in place of the RGB channels. The proposed deep unfolding network successfully combines the geometry of the PAN image with the spectral information of the HS data, while all other results are either affected by color artifacts like the MHFnet in the second row or do not detect accurately some relevant geometrical details like DiCNN and MSDCNN in the last row starting from top to bottom.