

# Middle-Output Deep Image Prior for Blind Hyperspectral and Multispectral Image Fusion

Jorge Bacca<sup>a</sup>, Christian Arcos<sup>a</sup>, Juan Marcos Ramírez<sup>b</sup> and Henry Arguello<sup>a</sup>

<sup>a</sup>Department of Computer Science, Universidad Industrial de Santander, Bucaramanga, 680002, Colombia

<sup>b</sup>IMDEA Networks Institute, Leganés, Madrid, 28918, Spain

## ARTICLE INFO

### Keywords:

Spectral image fusion  
unsupervised image fusion  
Deep image prior  
Deep Learning  
Learning degradation models

## ABSTRACT

Spectral image fusion combines low-spatial-resolution hyperspectral (HS) and low-spectral-resolution multispectral (MS) images to estimate a high-resolution (HR) spectral image. Although recent fusion techniques based on supervised deep learning have shown promising results, these methods require large training datasets involving expensive acquisition costs and long training times. In contrast, unsupervised HS and MS image fusion based on the deep image prior (DIP) methodology offers the advantage of adaptability to images with different distributions. However, existing unsupervised methods rely on the assumption of linear degradation models and require precise knowledge of these models for optimal performance. To overcome these challenges, we propose the Middle-Output Deep Image Prior (MODIP) for unsupervised blind HS and MS image fusion. MODIP is based on the DIP model and produces the fused image at an intermediate layer within the network. The architecture comprises two upscaling convolutional generators that reconstruct the HR spectral image from HS and MS inputs, along with two networks that appropriately downscale the estimated HR image to match the available MS and HS datasets, learning the non-linear degradation models. The network parameters of MODIP are jointly and iteratively adjusted by minimizing a proposed composite loss function. Importantly, this approach can handle scenarios where the degradation operators are unknown or partially estimated. To evaluate the performance of MODIP, we test the fusion approach on two simulated spectral image datasets (Pavia University and Salinas Valley) and a real dataset obtained through a testbed implementation in an optical lab. Extensive simulations demonstrate that MODIP outperforms other unsupervised model-based image fusion methods.


## 1. Introduction

Hyperspectral (HS) imaging has become an analysis tool that exploits the benefits of both conventional imaging and spectroscopy to capture information of bi-dimensional (2D) scenes across hundreds of spectralbands Shaw and Burke (2003), Ghamisi et al. (2017). The detailed spectral information provided by an HS image allows the identification of different materials in the scene, which has boosted its use in several applications such as precision agriculture Dale et al. (2013), remote sensing Bioucas-Dias et al. (2013), biomedical imaging Vo-Dinh et al. (2003), among others.

According to current technology, HS images typically show low-spatial resolution to achieve the desired signal-to-noise ratio (SNR). On the other hand, multispectral (MS) images are data sets that obtain detailed spatial information about the scenes but with lower spectral-resolution compared to that exhibited by HS images. In this regard, HS and MS image fusion focuses on merging the relevant information in HS and MS images to obtain a high-resolution spectral image as it would have been obtained with a single specialized camera Yokoya et al. (2017). In the last two decades, various methods have been developed to address the fusion problem based on different approaches, including multi-resolution analysis Liu (2000), Aiazzi et al. (2006), component substitution Aiazzi et al. (2007), spectral

unmixing Yokoya et al. (2012), and Bayesian estimation Wei et al. (2015). These methods characterize the observed HS and MS images as degraded versions of the target high-resolution spectral image. In consequence, these model-based techniques develop their procedures based on both the knowledge of the degradation operators describing observations or approximation of its, and the formulation of an optimization problem that usually is solved by an iterative algorithm.

Since the outstanding performance achieved by convolutional neural networks (CNNs) in computer vision and image processing tasks, supervised deep learning fusion techniques have been recently proposed. For example, in Xie et al. (2022) the authors provide an Unet network based on 1D convolution on the spectrum or in Palsson et al. (2017), Lai et al. (2017), Wang et al. (2022) three-dimensional (3D) convolutional neural network is used to fuse HS and MS images. On the other hand, deep algorithm unrolling techniques have been recently developed to solve the fusion problem Xie et al. (2019, 2020), Ramirez et al. (2021), Jacome et al. (2021). A method based on the subspace representation and a CNN-based grayscale image denoiser is developed in Dian et al. (2020). Moreover, a spatial-spectral reconstruction network is reported in Zhang et al. (2020) to perform HS and MS image fusion, and recently Transformer networks with attention module are proposed Qiao et al. (2023), Jia et al. (2023), Wang et al. (2023). These techniques typically require large training data sets to optimize the network parameters, and the performance is tailored to the subspace spanned by the training data.

 jorge.bacca1@correo.uis.edu.co (J. Bacca)

ORCID(s): 0000-0001-5264-7891 (J. Bacca); 0000-0002-6894-5591 (C.

Arcos); 0000-0003-0000-1073 (J.M. Ramirez); 0000-0002-2202-253X (H. Arguello)

Recently, the unsupervised deep image prior (DIP) scheme introduced in Ulyanov et al. (2018) overcomes these limitations under the premise that some deep learning architectures can capture the image statistics prior in the network weights without using a training set. This is, an image recovery problem is mapped into a network-based image generation architecture that requires the observation images and the exact knowledge of the degradation operators. This method has also been used in some inverse problems such as single image super-resolution Sidorov and Hardeberg (2019), denoising Sidorov and Hardeberg (2019), compressive spectral imaging (CSI) reconstruction Bacca et al. (2021), Gelvez et al. (2021) and recently in the fusion problem Wang et al. (2023), Zheng et al. (2020), Liu et al. (2020), Wang et al. (2020), Sun et al. (2021), Li et al. (2023). However, to obtain good reconstructions, the DIP method needs to know the exact degradation models because the desired image is obtained by minimizing the  $\ell_2$ -norm between the observations and degraded versions of the recovered image. In the context of spectral fusion, it is crucial to emphasize that in real-world environments, the degradation operators are typically unknown or partially estimated. This issue arises due to various factors, including calibration intricacies, temporal inconsistencies, spatial misalignments, and spectral variability Vivone (2023). Consequently, these complexities significantly complicate the estimation process of these operators.

The methodologies that tackle the problem of estimating degradation operators and performing fusion are referred to as blind hyperspectral image fusion. Authors in Gu et al. (2019) have employed datasets to learn the degradation models, which are then incorporated into the fusion scheme. Additionally, the DIP methodology have modify to consider blind hyper-spectral image fusion Liu et al. (2021), Nie et al. (2020), Bandara et al. (2021). For instance, Bandara et al. (2021) focuses on learning a linear degradation model for the panchromatic (PAN) image. In Nie et al. (2020), two linear operators are estimated, one for MS, and another for HS images, which are estimated alternatively with the fusion image reconstruction. And in Liu et al. (2021) learn jointly the linear degradation models and the weights of the fusion network in an end-to-end (E2E) scheme Arguello et al. (2023). Nevertheless, it is important to note that in real-world scenarios, the linear assumption in the degradation operators may not accurately capture the complex relationship between the high-resolution image and the MS and HS images Gao et al. (2021), Hong et al. (2018).

Therefore, this paper proposes Middle-Output Deep Image Prior (MODIP) for unsupervised blind HS and MS image fusion. In particular, the proposed architecture includes DIP instances as processing blocks of a connected network that iteratively learns, in an unsupervised fashion, both the upscaling non-linear mapping that generates the fused image and the degradation operators describing the HS and MS images. More precisely, the DIP instances are compactly expressed in a single model, called MODIP, where the entire weight set of the connected network is continuously

optimized by minimizing a single loss function using an end-to-end framework. In this approach, the fused image is obtained at an intermediate stage of the connected system. Furthermore, the DIP processing blocks included in MODIP consist of unbalanced networks whose inputs and outputs exhibit different data resolutions. Importantly, this approach can handle scenarios where the degradation operators are unknown or partially estimated. The performance of the proposed architecture was experimentally evaluated on two simulated data sets: Pavia University and Salinas Valley and a real dataset obtained through a testbed implementation in an optical lab. We also show the remarkable performance of the proposed approach compared to those obtained by other unsupervised methods. Specifically, the contributions of this work are summarized as follows.

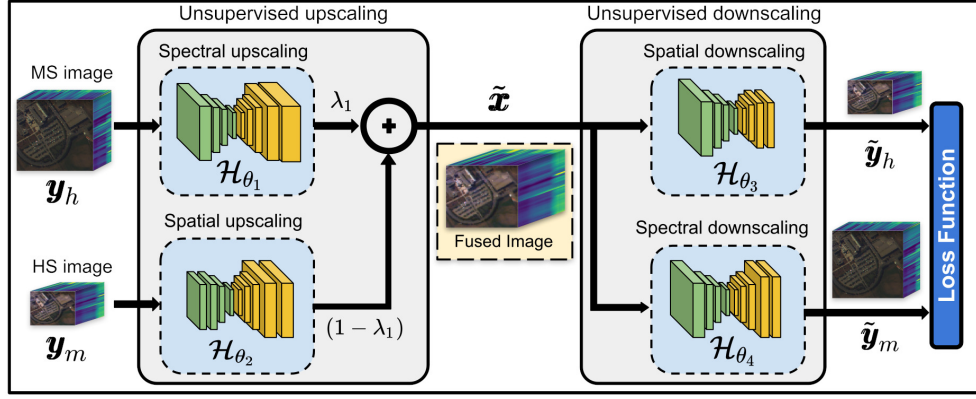
1. We propose the MODIP for unsupervised blind HS and MS image fusion. This architecture includes multiple DIP model instances as processing blocks, whose weights are jointly optimized by minimizing a single loss function. In essence, the proposed architecture iteratively optimizes the network weights and versions of the observed (HS and MS) images. Moreover, the fused image is obtained at an intermediate stage of the proposed architecture.
2. To adapt the DIP model instances to the fusion problem, we describe convolutional generators with different input and output resolutions that learn both the upscaling function generating the fused image and degradation operators describing the input images.
3. We also propose to estimate the degradation operators through a non-linear model learned in an end-to-end manner in the MODIP.
4. The MODIP is evaluated in a real-test-bed implementation validation the assumption of the learned non-linear propagation models.

This paper is organized as follows. Section 2 briefly describes the background, and the proposed architecture is presented in Section 3. Section 4 details the architecture settings and Section 5 displays the simulation results. Section 8 synthesizes some concluding remarks and future work.

## 2. Preliminaries

### 2.1. Spectral image fusion

Spectral image fusion attempts to estimate a high-spatial and spectral-resolution image by combining the information captured by both a high-spectral-resolution hyperspectral (HS) image and a high-spatial-resolution multispectral (MS) image. To this end, let  $\mathbf{x} \in \mathbb{R}^{MNL}$  be the ideal high-spatial and spectral-resolution image in a vector form with  $M \times N$  pixels and  $L$  spectral bands. On the other hand, consider  $\mathbf{y}_h \in \mathbb{R}^{M'N'L}$  the HS image whose spatial resolution is  $M' \times N'$ , with  $M' = M/p$ ,  $N' = N/p$ , where  $p$  is the spatial decimation factor. Furthermore,  $\mathbf{y}_m \in \mathbb{R}^{MN'L}$  denotes the MS image with  $L' = L/q$  as the decimated spectral bands where  $q$  is the spectral downsampling factor. In general, the



**Figure 1:** Schematic of the proposed MODIP architecture for unsupervised HS and MS image fusion. MODIP receives the MS and HS images as inputs. Each input image passes by an upsampling convolutional network that attempts to yield high-spatial- and-spectral-resolution features. Then, an affine combination is applied to obtain the fused spectral image. Afterward, the fused spectral image goes through downsampling networks to estimate approximated versions of the HS and MS images. Notice that all network parameters are jointly optimized by minimizing a single loss function.

acquisition models for HS and MS images are given by

$$\mathbf{y}_h = \mathbf{D}_h \mathbf{x} + \boldsymbol{\eta}_h, \quad (1)$$

$$\mathbf{y}_m = \mathbf{D}_m \mathbf{x} + \boldsymbol{\eta}_m, \quad (2)$$

where  $\mathbf{D}_h \in \mathbb{R}^{M'N' \times L \times MNL}$  denotes the spatial decimation operator,  $\mathbf{D}_m \in \mathbb{R}^{MNL' \times MNL}$  is the spectral downsampling operator characterizing the multispectral acquisition system,  $\boldsymbol{\eta}_h$  and  $\boldsymbol{\eta}_m$  represent the additive noise vectors that describe the contamination affecting HS and MS measurement systems, respectively Rasti et al. (2017), Wei et al. (2015). The noise vector entries are typically described as independent and identically distributed (iid) random samples obeying a Gaussian distribution.

When decimation operators are known, the minimization of the sum of squared errors between the low-resolution images and degraded versions of the fused image can be formulated as the performance criterion to be optimized Bacca et al. (2017). However, this is an ill-posed inverse problem whose estimations frequently induce blur effects and artifacts. From the Bayesian point of view, the fusion problem can be redefined as a maximum *a posteriori* probability (MAP) estimation. More precisely, this approach focuses on solving the optimization

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \frac{1}{2} \|\mathbf{y}_m - \mathbf{D}_m \mathbf{x}\|_2^2 + \frac{1}{2} \|\mathbf{y}_h - \mathbf{D}_h \mathbf{x}\|_2^2 + \lambda P_{\mathbf{x}}(\mathbf{x}) \quad (3)$$

where  $P_{\mathbf{x}}(\mathbf{x})$  is the prior term that statistically describes previously known information of the target image with the goal of recovering scene details lost in the degradation process, and  $\lambda$  is the regularization parameter that controls the trade-off between the sum of squared errors and the penalty term. Notice that image priors have been extensively used to solve the data fusion problem from HS and MS images, including the Tikhonov regularization Ramirez and Arguello (2019), the sparsity-inducing term Akhtar et al. (2014), Wei et al. (2015), the total variation (TV) norm Simoes et al. (2014) and low-rank promoting norms Rasti et al. (2017), Bacca et al. (2019).

## 2.2. Deep image prior

The deep image prior (DIP) is a network-based framework introduced by Ulyanov et al. (2018) to solve imaging inverse problems in an unsupervised way. This framework assumes that the structure of a convolutional generator can capture the statistic priors that characterize image details without resorting to a learning stage. More precisely, to obtain an estimate of the desired image, the DIP approach fits the parameters of a convolutional generator network minimizing a loss function that considers the image observation model and the available degraded image. In the context of spectral image fusion, the DIP approach can be adapted to obtain the high-resolution image as follows

$$\begin{aligned} \hat{\boldsymbol{\theta}}_1 &= \arg \min_{\boldsymbol{\theta}_1} \|\mathbf{y}_h - \mathbf{D}_h \mathcal{H}_{\boldsymbol{\theta}_1}(\mathbf{z})\|_2^2 + \|\mathbf{y}_m - \mathbf{D}_m \mathcal{H}_{\boldsymbol{\theta}_1}(\mathbf{z})\|_2^2, \\ \hat{\mathbf{x}} &= \mathcal{H}_{\hat{\boldsymbol{\theta}}_1}(\mathbf{z}), \end{aligned} \quad (4)$$

with  $\mathbf{z}$  as a fixed random noise Liu et al. (2021, 2020) or either the MS and HS Wang et al. (2023), Gao et al. (2021), where  $\mathcal{H}_{\boldsymbol{\theta}_1}(\cdot)$  stands for the convolutional generator network receiving the MS image only and returning the fused image, and  $\boldsymbol{\theta}_1$  represents the network parameters. Note that the loss function minimization considers the high-spectral resolution information embedded in the HS and MS images. In general, DIP-based methods optimize network parameters  $\boldsymbol{\theta}_1$  by running a conventional deep learning optimization algorithm such as gradient descent from scratch (random initialization). In contrast to traditional deep learning approaches that train the network structures using large databases, the DIP approach estimates the network weights  $\boldsymbol{\theta}_1$  using the available images ( $\mathbf{y}_h, \mathbf{y}_m$ ) and the exact knowledge of the downsampling operators ( $\mathbf{D}_h, \mathbf{D}_m$ ) describing the measurements process. Afterward, at every iteration, an estimation of the target image  $\mathbf{x}$  can be obtained from the latent input  $\mathbf{z}$  by implementing the trained image generator network, i.e.,  $\hat{\mathbf{x}} = \mathcal{H}_{\hat{\boldsymbol{\theta}}_1}(\mathbf{z})$ .

In general, good performance with the DIP approach requires correct knowledge of the downsampling operators,

so DIP does not exploit the generator power provided by convolutional networks to learn the downgrade operators introduced by the measurement process. Therefore, a spectral image fusion framework is required that considers both the remarkable performance of the unsupervised DIP model and the power of the convolutional networks to describe the image sampling operators.

### 3. Proposed approach

In this section, we introduce the middle-output deep image prior (MODIP). First, we present an overview of the proposed deep learning architecture for blind unsupervised spectral image fusion, i.e., the proposed method does not assume knowledge of the degradation operators. Afterward, we analyze the loss function that iteratively optimizes the network parameters. Finally, we include a conceptual visualization of the proposed approach using a manifold representation.

#### 3.1. Architecture overview

A schematic representation of the proposed MODIP architecture for blind unsupervised HS and MS image fusion is shown in Fig. 1. Notice that the DIP model, described in the previous section, receives a single input latent vector and generates the high-resolution image  $\tilde{\mathbf{x}}$  by minimizing a loss function that considers the observed images and the degradation operator describing measurements. In contrast to the DIP approach, MODIP receives two input latent vectors with different sizes (size related to the HS image and the MS image) to exploit the high-resolution information embedded in both observed images. Instead of a random fixed vector, our input latent vectors are corrupted versions of the HS ( $\mathbf{z}_h$ ) and MS ( $\mathbf{z}_m$ ) images, respectively. As can be seen in Fig. 1, each input image goes through a particular convolutional generator network with the goal of estimating the fused image.

On one hand, the MS image goes through a convolutional neural network (CNN) whose mapping can be written as  $\tilde{\mathbf{x}}' = \mathcal{H}_{\theta_1}(\mathbf{z}_m)$ , where  $\mathcal{H}_{\theta_1}(\cdot) : \mathbb{R}^{M'N'L} \rightarrow \mathbb{R}^{MNL}$  stands the CNN spectral upscaling that obtains a high-resolution spectral image from the MS image. On the other hand, the HS image passes by a different CNN whose mapping is denoted as  $\tilde{\mathbf{x}}'' = \mathcal{H}_{\theta_2}(\mathbf{z}_h)$ , with  $\mathcal{H}_{\theta_2}(\cdot) : \mathbb{R}^{M'N'L} \rightarrow \mathbb{R}^{MNL}$  representing the CNN spatial upscaling yielding a high-resolution spectral image from the HS image. Subsequently, the fused image is estimated through an affine combination that considers the high-resolution outputs yielded by the upscaling CNNs. More precisely, the fusion operation can be described as

$$\begin{aligned}\tilde{\mathbf{x}} &= \lambda_1 \tilde{\mathbf{x}}' + (1 - \lambda_1) \tilde{\mathbf{x}}'' \\ &= \lambda_1 \mathcal{H}_{\theta_1}(\mathbf{z}_m) + (1 - \lambda_1) \mathcal{H}_{\theta_2}(\mathbf{z}_h),\end{aligned}\quad (5)$$

where  $\lambda_1$  is the coefficient of the two-term affine combination. For the sake of simplicity, we represent the image fusion stage using a single network as

$$\tilde{\mathbf{x}} = \mathcal{G}_\vartheta(\mathbf{y}_m; \mathbf{y}_h), \quad (6)$$

where  $\vartheta = \{\theta_1, \theta_2, \lambda_1\}$ .

Afterward, a convolutional network represented by  $\mathcal{H}_{\theta_3}(\cdot) : \mathbb{R}^{MNL} \rightarrow \mathbb{R}^{M'N'L}$  performs a downscaling operation on the fused image estimate  $\tilde{\mathbf{x}}$  to recover an approximated version of the HS image. In essence, this network unsupervisedly learns the degradation function describing the HS image, and its output can be depicted as  $\tilde{\mathbf{y}}_h = \mathcal{H}_{\theta_3}(\tilde{\mathbf{x}})$ . Similarly, the fused image goes through a CNN  $\mathcal{H}_{\theta_4}(\cdot) : \mathbb{R}^{MNL} \rightarrow \mathbb{R}^{MNL}$  that yields an estimation of the MS image, i.e.  $\tilde{\mathbf{y}}_m = \mathcal{H}_{\theta_4}(\tilde{\mathbf{x}})$ . As can be observed in Fig. 1, the proposed architecture comprises an unsupervised upscaling stage followed by an unsupervised downscaling stage. Therefore, the proposed network meets two goals, performing an unsupervised spectral image fusion (which is obtained at the middle of the MODIP and learning the downscaling functions describing observed HS and MS images.

#### 3.2. Image fusion based on the DIP approach

In this work, we exploit the benefits of the DIP approach to solve the image fusion problem. In this regard, consider  $\theta = \{\vartheta \cup \theta_3 \cup \theta_4\}$  as the set of all learnable parameters of the MODIP architecture. Therefore, the weights of the entire architecture are jointly adjusted by minimizing a single loss function of the form

$$\begin{aligned}\hat{\theta} &= \arg \min_{\theta} \left\{ \tau_1 \left\| \mathbf{y}_h - \mathbf{D}_h \mathcal{G}_\vartheta(\mathbf{z}_m; \mathbf{z}_z) \right\|_2^2 \right. \\ &\quad + \tau_2 \left\| \mathbf{y}_m - \mathbf{D}_m \mathcal{G}_\vartheta(\mathbf{z}_m; \mathbf{z}_z) \right\|_2^2 \\ &\quad + \lambda_2 \left\| \mathbf{y}_h - \mathcal{H}_{\theta_3}(\mathcal{G}_\vartheta(\mathbf{z}_m; \mathbf{z}_h)) \right\|_2^2 \\ &\quad \left. + (1 - \lambda_2) \left\| \mathbf{y}_m - \mathcal{H}_{\theta_4}(\mathcal{G}_\vartheta(\mathbf{z}_m; \mathbf{z}_h)) \right\|_2^2 \right\},\end{aligned}\quad (7)$$

where  $\hat{\theta} = \{\hat{\vartheta} \cup \hat{\theta}_3 \cup \hat{\theta}_4\}$  is the entire set of parameters of the MODIP architecture,  $\lambda_2 > 0$  is a penalty parameter that controls the influence of the downscaling network, and  $(\tau_1, \tau_2) \in \{0, 1\}$  are indicator variables denoting the knowledge or not of the degradation operators. Assuming that upscaling networks fetch the prior spatial-spectral information of the high-resolution image, the fused image is obtained at the intermediate stage of the MODIP architecture, in other words

$$\tilde{\mathbf{x}} = \mathcal{G}_{\hat{\vartheta}}(\mathbf{z}_m; \mathbf{z}_h), \quad (8)$$

i.e., once (7) is solved, the optimal weights ( $\hat{\vartheta}$ ) of the MODIP produce the fused spectral image. Furthermore, to learn the degradation operators describing the available data, estimations of the HS and MS images are obtained at the output of the downscaling networks, i.e.

$$\tilde{\mathbf{y}}_h = \mathcal{H}_{\hat{\theta}_3}(\tilde{\mathbf{x}}) = \mathcal{H}_{\hat{\theta}_3}(\mathcal{G}_{\hat{\vartheta}}(\mathbf{z}_m; \mathbf{z}_h)), \quad (9)$$

$$\tilde{\mathbf{y}}_m = \mathcal{H}_{\hat{\theta}_4}(\tilde{\mathbf{x}}) = \mathcal{H}_{\hat{\theta}_4}(\mathcal{G}_{\hat{\vartheta}}(\mathbf{z}_m; \mathbf{z}_h)). \quad (10)$$

In summary, the proposed approach addresses the image fusion problem using the DIP approach. More precisely, MS

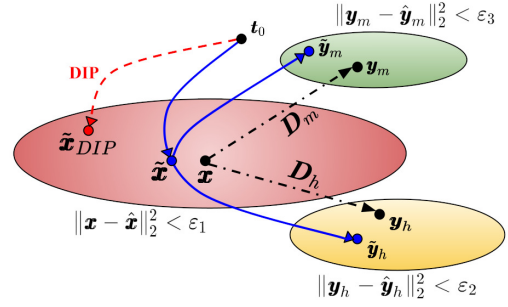
and HS images feed the MODIP architecture, and the set of network parameters are randomly initialized. Then, the proposed method optimizes all network weights by minimizing the composed loss function (7).

Notice that the loss function includes the sum of squared errors that is commonly used to solve the spectral image fusion problem. Specifically, this sum requires the low-resolution images ( $\mathbf{y}_h, \mathbf{y}_m$ ) and the degradation operators ( $\mathbf{D}_h, \mathbf{D}_m$ ). Furthermore, two explicit regularization terms are included in the loss function that penalize the capacity of the proposed cascaded architecture to learn the degradation operators describing the HS and MS images. The purpose of these regularizers is to improve the performance of the upscaling network such that the fused image correctly downscales to the available observations in an unsupervised fashion. It can be observed in (7) that additional penalty terms enable the downscaling network to learn the degradation functions. Thus, the downscaling network can be seized when the image acquisition models are not accurately known.

Finally, Fig. 2 illustrates a visualization of the image space for the fusion problem using the proposed MODIP architecture. As can be seen in this figure, the image of interest  $\mathbf{x}$  lies in the manifold surface with low energy errors, i.e.  $\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 < \varepsilon_1$ , whose representation is in red. Notice that the low-resolution images  $\mathbf{y}_m$  and  $\mathbf{y}_h$  can be obtained directly by applying, respectively, the degradation operators  $\mathbf{D}_m$  and  $\mathbf{D}_h$  to the high-resolution image  $\mathbf{x}$ . In addition, the HS image  $\mathbf{y}_h$  belongs to a low-resolution manifold of points with low energy errors  $\|\mathbf{y}_h - \hat{\mathbf{y}}_h\|_2^2 < \varepsilon_2$  that is shown in yellow. Similarly, the MS image belongs to the manifold area illustrated in green  $\|\mathbf{y}_m - \hat{\mathbf{y}}_m\|_2^2 < \varepsilon_3$ . The deep image prior (DIP) approach optimizes the network weights such that the image estimate  $\tilde{\mathbf{x}}_{DIP}$  belongs to the manifold surface with low energy errors. This approach typically exhibits better results compared to those obtained by the optimization-based approaches. In contrast to the DIP technique, the proposed approach optimizes the deep network parameters to obtain the fused image estimate that correctly decimates HS and MS images. The minimization of the end-to-end loss function aims at obtaining low-resolution images that reach the lower-dimensional manifolds with low measurement errors.

## 4. Network Settings

The proposed image fusion framework is based on four unbalanced CNNs  $\mathcal{H}_{\theta_1}, \mathcal{H}_{\theta_2}, \mathcal{H}_{\theta_3}$  and  $\mathcal{H}_{\theta_4}$  which are explained in Section 4.0.1. The term unbalanced refers to the fact that the inputs and the outputs of the autoencoders share different dimensions. In essence, these convolutional generators receive a spectral image with a specific dimension and produce an output image with a different spatial or spectral dimension. Notice that any subnetwork that meets the input-and output dimension criteria can be used for the unbalanced networks, however, we desired to use Skip-Net and Unet-based according to the results shown in Ulyanov et al. (2018), Bacca et al. (2021). Experimentally, we set



**Figure 2:** Image space representation of the effects induced by the MODIP architecture. The image of interest  $\mathbf{x}$  lies in the manifold surface with low energy where the MS and HS can be obtained by applying the trained operators.

the latent input images are the HS and MS corrupted with Gaussian noise of 15 dB of SNR. Although the best results can be achieved by carefully tuning an architecture, the chosen networks work well in practice without many hyperparameter settings.

### 4.0.1. Unbalanced Networks

The networks used are based on encoder-decoder networks, since receiving images with different inputs and outputs shapes. For  $\mathcal{H}_{\theta_1}$ , which receives the MS image and obtains features with the spatial dimension  $M \times N \times L$  we used a Skip-Net-based network. This consists of  $B$  successive down-sampling and up-sampling blocks; each performs a convolution, batch normalization, and applies a LeakyReLU activation function. All the convolutional filters are  $3 \times 3$ , with a stride of 2 during downsampling, and for the upsampling model of the skip-Net, we used the bilinear function. The number of features in the last layer increases according to the decimation factor to meet the spectral dimensions, so we learn a hierarchy of features to obtain the high-resolution image in the middle of the system.

On the other hand, for  $\mathcal{H}_{\theta_2}$  which receives the HS image, we used an Unet-based network Bacca et al. (2021). Unet is an hourglass architecture with skip connections composed of a convolution operator with  $3 \times 3$  filters followed by a LeakyReLU activation and Maxpooling with 2. For the upsampling part, the nearest interpolation and the features are concatenated between the encoder and decoder at the same depth level. To obtain the desired image, an additional upsampling block is included in the unbalance Unet. Finally,  $\mathcal{H}_{\theta_3}$  and  $\mathcal{H}_{\theta_4}$  in charge of learning the degradation models of the HS and MS images, we used an Unet-based and Skip-Net-based, respectively. For  $\mathcal{H}_{\theta_3}$  we eliminate the last upsampling layer according to the decimation factor  $p$  and for the  $\mathcal{H}_{\theta_4}$  we reduce the number of filters in the last layer to  $L'$  to meet the size criteria.

## 5. Simulations results

This section assesses the performance of the proposed image fusion approach using two hyperspectral image datasets: Pavia University and Salinas Valley. First, we test the

proposed network under different parameter settings. Then, we compare the performance of the proposed image fusion method with respect to those obtained by other unsupervised approaches using three image quality metrics: peak signal-to-noise ratio (PSNR) spectral angle mapper (SAM), and structural similarity (SSIM) index Wang et al. (2004). The results obtained with the MODIP architecture are obtained using 50000 updating steps and a learning rate of 0.001. It is worth highlighting that our method does not need training data; therefore, the network weights are randomly initialized using the Xavier method Narkhede et al. (2022), and weights are subsequently updated using the Adam gradient descent optimization Kingma and Ba (2014). The MODIP is implemented in Python using the Pytorch library. The source code of the proposed image fusion method can be downloaded from [https://github.com/TottiPuc/MODIP\\_superResolution](https://github.com/TottiPuc/MODIP_superResolution).

## 5.1. Datasets

### 5.1.1. Salinas Valley

The AVIRIS sensor acquired this dataset over a set of crops in the Valley of Salinas, USA Jet Propulsion Laboratory, NASA (2019). More precisely, this image exhibits a spatial resolution of 3.7 meters per pixel with  $217 \times 512$  pixels and 224 spectral bands in the wavelength interval from 400nm to 2500nm.

### 5.1.2. Pavia University

This spectral image was captured by the Reflective Optics System Imaging Spectrometer (ROSIS-03) over an urban area of the University of Pavia, Italy de Inteligencia Computacional (2008). Specifically, this dataset exhibits a spatial resolution of 1.3 meters per pixel and contains  $610 \times 340$  pixels and 103 spectral bands in the wavelength range from 430nm to 860nm. This work uses a subset of 96 spectral bands by removing the spectral bands within the wavelength range 830-860nm.

### 5.1.3. Cave Dataset

This data set contains 32 indoor HSIs with 31 visible spectral bands from 400nm to 700nm with 10nm between them and  $512 \times 512$  of spatial resolution. To compare with state-of-the-art methods, we use the same training and testing partition suggested by Nie et al. (2020), Shi et al. (2023), Hu et al. (2021) where 20 are used for training and the other 12 images used for testing. It is important to highlight that the training is only used for state-of-the-art data-driven methods, not the proposed method, since our MODIP is an unsupervised strategy.

## 5.2. HS and MS image degradation models

This work obtains HS and MS images by degrading the corresponding high-resolution datacube across spatial and spectral coordinates, respectively. More precisely, for the HS images, we first filter each spectral band of the high-resolution image using a  $3 \times 3$  Lanczos kernel. Subsequently, we apply a spatial downsampling operator to every filtered band using a  $p^2 : 1$  decimation ratio with  $p$  as the spatial

**Table 1**

MODIP performance for different hyperparameter settings for the Salinas Dataset. Downsampling factors fixed to  $p = 4$  and  $q = 4$ . Input images contaminated with AWGN noise at SNR = 30 dB. The setting  $(\tau_1 = 0, \tau_2 = 0)$  indicates that we do not possess knowledge about degradation operators.

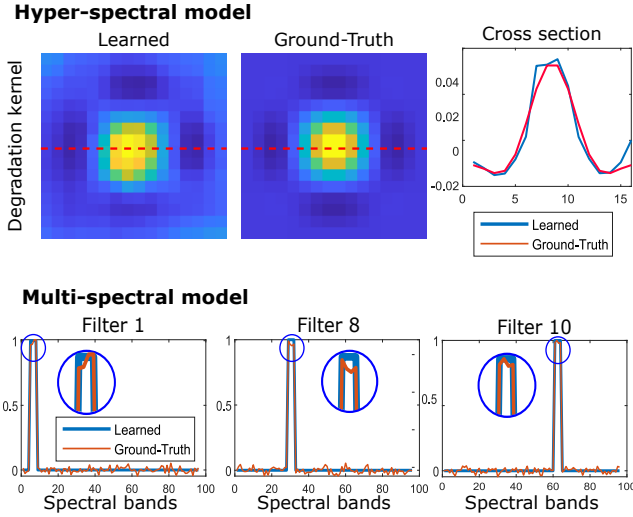
$\lambda_1$	$\lambda_2$	$\tau_1$	$\tau_2$	PSNR $\uparrow$	SAM $\downarrow$	SSIM $\uparrow$
1	1	1	1	44.35	2.04	0.995
1	1	0	0	44.96	2.53	0.995
0	0	1	1	37.03	2.28	<u>0.996</u>
0	0	0	0	41.76	1.96	0.993
1	0	1	1	45.92	2.08	0.991
1	0	0	0	44.19	2.33	0.995
0	1	1	1	35.82	2.41	0.995
0	1	0	0	40.81	<b>1.08</b>	0.993
0.5	0.5	1	1	<b>47.89</b>	1.79	<b>0.996</b>
0.5	0.5	0	0	<u>45.99</u>	<u>1.45</u>	0.995

decimation factor. On the other hand, the MS images are synthesized by downsampling the high-resolution image along the spectral axis. In this case, every spectral band of the MS image is obtained by averaging  $q$  nonoverlapping spectral bands of the high-resolution image along the spectral axis with  $q$  as the spectral decimation factor. To consider the measurement errors induced during the sensing process, HS and MS images are contaminated using the additive Gaussian white noise (AWGN) model with SNR fixed to 30 dB.

## 5.3. Characterization of MODIP

The first experiment illustrates the effect of varying the MODIP hyperparameters. Notice that  $\lambda_1$  in Eq.(5) controls the influence of the input (MS and HS) images on the fused output obtained in the middle stage of the proposed architecture. Specifically, if  $\lambda_1 = 0$ , the fused image only considers the HS image. On the contrary, if  $\lambda_1 = 1$ , the high-resolution image only considers the MS image. Similarly,  $\lambda_2$  in Eq.(7) controls the effect of degradation networks on the loss function minimization. In this case, if  $\lambda_2 = 0$ , the loss function discards the spectral downsampling network, whereas the loss function discards the spatial downsampling branch when  $\lambda_2 = 1$ . Two essential and exciting hyperparameters of the proposed model are  $\tau_1$  and  $\tau_2$  that consist of indicator variables denoting the knowledge level of the degradation operators, i.e. if  $\tau_1 = 0$ , the loss function discards the term that includes the degradation operator describing the HS image  $\mathbf{D}_h$ . Therefore, we can set  $\tau_1 = 0$  when  $\mathbf{D}_h$  is unknown. On the other hand, we can fix  $\tau_2 = 0$  when the degradation operator  $\mathbf{D}_m$  is unknown. Furthermore, we can set  $\tau_1 = \tau_2 = 0$  when both degradation operators are unknown.

We use the Salinas Valley dataset to evaluate the proposed fusion approach for different hyperparameter settings. For this experiment, the degradation factors are fixed to  $p = 4$  and  $q = 4$  to generate the HS image and MS image, respectively. Table 1 summarizes the quantitative results obtained by varying the MODIP hyperparameters, where the best results are in bold and the second-best values are underlined.



**Figure 3:** Visual representation of the degradation operators. (Top) The learned degradation kernel and its cross-section. (Bottom), the learned spectral filters, and the ground-truth.

It can be seen that the best PSNR values are obtained when the MODIP considers the information provided by both the MS image and HS image corresponding to the case  $\lambda_1 = 0.5$ . Indeed, the best PSNR values is yielded when considering the two degradation networks ( $\lambda_2 = 0.5$ ) and the degradation models are known or available ( $\tau_1 = 1, \tau_2 = 1$ ). As expected, the proposed fusion approach with the knowledge of the degradation operators shows an improvement of at least 2dB compared to the performance when the downsampling models are not known. However, it should be noted that, when the degradation models are not known, the proposed generator architecture exhibits outstanding performance in terms of PSNR, SAM, and SSIM.

#### 5.4. Learning the degradation models

One of the main contributions of the proposed fusion method is that is not necessary to know the exact degradation models for the MSI and the HSI, since these operators are learned using the subnets  $\mathcal{H}_{\hat{\theta}_3}$  and  $\mathcal{H}_{\hat{\theta}_4}$ . To show the effectiveness of the proposed method, we run the previous experiment with  $\tau_1, \tau_2 = 0$ , i.e., which discards the knowledge embedded in  $\mathbf{D}_h$  and  $\mathbf{D}_m$ . To obtain the multi-spectral kernel, a process to obtain the full-resolution spatial point spread function of the  $\mathcal{H}_{\hat{\theta}_3}$  is performed Arguello et al. (2021). This process consists of passing white spectral points individually through the  $\mathcal{H}_{\hat{\theta}_3}$  operator and recording the measurements. This light response is performed through a spatial window, and the central value of the measurements is arranged in a matrix denominated as the kernel. Similarly, a white image for each spectral bandpass through  $\mathcal{H}_{\hat{\theta}_4}$  and obtain the intensity for each filter. Fig.3 shows the estimated and the original degradation model for the kernel in the MS image and the filters in the HS image. It can be seen that although the models are entirely unknown, the proposed network can estimate the degradation models correctly, maintaining the accuracy of the model.

**Table 2**

Mean quantitative comparison results of the fused spectral image for the CAVE testing dataset. The best performance of each experiment is shown in bold, and the second-best is underlined.

Method	PSNR $\uparrow$	SAM $\downarrow$	SSIM $\uparrow$
PANet Yang et al. (2017)	30.68	13.68	0.864
HSRnetHu et al. (2021)	41.05	7.94	0.975
SSRnetZhang et al. (2020)	41.73	8.17	0.975
MHFnet Xie et al. (2019)	36.46	22.12	0.951
EDBIN Wang et al. (2021)	40.68	8.96	0.969
DDPM-Fus Shi et al. (2023)	<u>43.66</u>	<u>5.69</u>	<u>0.986</u>
<b>MODIP-our</b>	<b>49.19</b>	<b>3.33</b>	<b>0.995</b>

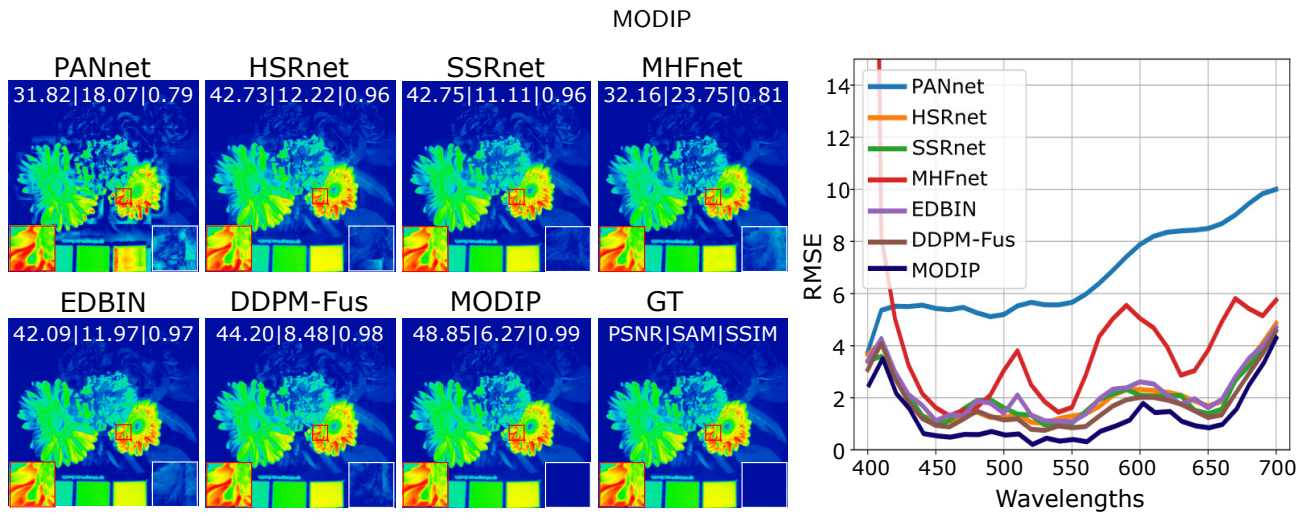
## 6. State-of-the-art Comparison

In this section, we compare the performance of the proposed network architecture with respect to other open-source code state-of-the-art methods, PANet Yang et al. (2017), HSRnet Hu et al. (2021), SSRnet Zhang et al. (2020), MHFnet Xie et al. (2019), EDBIN Wang et al. (2021) and DDPM-Fus Shi et al. (2023), where the last one presents the state-of-the-art in blind spectral image fusion. The evaluated factors are  $p = 8$  and  $q = 32$ . Each generated image is contaminated with additive white Gaussian noise at SNR = 30 dB. The parameters of the proposed network-based technique are set to  $\lambda_1 = 0.5, \lambda_2 = 0.5, \tau_1 = 1$ , and  $\tau_2 = 1$ . Table 2 shows the results obtained by the state-of-the-art methods and the proposed MODIP for the CAVE dataset. From the table, we can see that the MODIP method provides the best results in terms of image quality and is indeed effective in solving the fusion problem. This performance is obtained as the MODIP method introduces the idea of extracting multi-scale features, which makes full use of the advantages of each scale and integrates the degradation operator that describes the acquisition process capturing the HS image. That is, the MODIP method can still maintain a high performance compared with other single traditional even when spectral information is scarce.

In order to visualize some reconstruction, Figure 4 displays the fused high-resolution images in the 21st band (600 nm) obtained by the different fusion methods for a testing spectral image of the CAVE dataset were a zoom cropped region is highlighted with their corresponding residual map. Furthermore, the corresponding RMSE across the spectral band is illustrated. As can be seen from Fig. 4, the proposed approach exhibits the best image evidence of spatial quality in the recovery image and the zoomed version. In addition, they provide the lowest RMSE across the entire wavelength, which demonstrates the best spectral consistency.

## 7. Test-bed Implementation

An optical test bed was implemented in the HDSP Optics Laboratory at Universidad Industrial de Santander, Colombia, to carry out a real-data spectral fusion experiment and the data was provide in our previous work Gelvez-Barrera



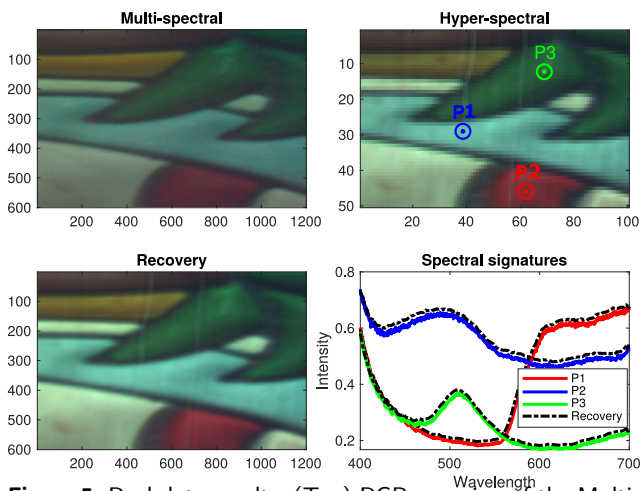
**Figure 4:** (Left) Image fusion results with their respective metrics in the 21st band (600 nm) of a testing spectral image of the CAVE dataset for the evaluated methods, where a zoom version and their corresponding residual map is highlighted. (Right) The corresponding RMSE along with spectral bands.

et al. (2023). The test-bed system is composed of an objective lens that focuses the scene onto the input plane of a 4f system with a beam-splitter located in the Fourier plane to split the light into an RGB commercial camera and a push-broom spectral camera. The SI arm is composed of an adjustable mechanical slit (Thorlabs VA100C-30 mm, 8-32 Tap). To obtain the spectral image, a relay lens located at 100 mm from a slit forms a parallel beam that reaches the 600 grooves/mm transmission grating that diffracts the light rays onto the sensor. The push-broom imagery spectrogram illuminates the slit with 1032 wavelengths in the spectral range 420 – 700nm with steps of 2.7nm. The SI acquisition performs 50 horizontal steps with a resolution of  $48\mu\text{m}$  so that the resulting SI has a spatial dimension of  $50 \times 100$  pixels and 1032 spectral bands. On the other hand, the Multispectral image is a crop region of  $600 \times 1200$  of the RGB camera. Consequently, the degradation factors are  $p = 12$  for the spatial decimation factor and  $q = 334$  for the

spectral, which are high degradation factors. Figure 5 (Top) shows an RGB visual representation of the acquired with the test-bed implementation and (Bottom) an RGB mapping of the reconstruction obtained with the MODIP. There it can be seen that the proposed method maintains the spatial structure of the image. Also, to see the spectral behavior, three spatial points are plotted of the recovery image, and the Hyperspectral image point signatures are used as reference. There it can be seen that the proposed method maintains the spectral signature in this real scenario.

## 8. Conclusions

This paper introduces the middle-output deep image prior (MODIP) architecture for blind unsupervised spectral image fusion from multispectral and hyperspectral images. Specifically, the proposed architecture relies on the deep image prior (DIP) approach assuming that the fused image prior statistics are fetched from the convolutional neural network structure. More precisely, MODIP includes up-scaling and down-scaling generator networks whose parameters are jointly optimized to estimate the fused image and the non-linear degradation models for the MS and HS image respectively. We tested the performance of the proposed deep architecture on two simulated spectral image datasets. First, we evaluated the performance of the proposed approach for different parameter settings and analyzed the behavior of the architecture when the information of the degradation operators is unknown or partially known. Furthermore, the results obtained by the MODIP outperformed those yielded by other unsupervised model-based techniques. Finally, the proposed fused method is evaluated with real data obtained in the HDSP group to validate the effectiveness of the MODIP.



**Figure 5:** Real-data results. (Top) RGB mapping of the Multi-spectral and Hyperspectral images. (Bottom) Visual (Bottom) RGB mapping of the recovery image with MODIP and some spectral signatures.

## Declaration of competing interest

The authors have no competing interests to declare.



## Acknowledgment

This paper was supported by the Vicerrectoría de Investigación y Extensión UIS, project code 3924 "Sistema óptico-computacional multiespectral de bajo costo en el infrarrojo cercano para la estimación de propiedades físico-químicas de granos secos de cacao mediante aprendizaje profundo". The work of Christian Arcos was supported by Minciencias through the postdoctoral research grant under code 848-2019. The work of Juan M. Ramirez has been supported by the Project ECID: Edge Computing for Intelligent Driving (PID2019-109805RB-I00) funded by the Spanish State Research Agency, Spanish Ministry of Science and Innovation.

## References

- G. A. Shaw, H. K. Burke, Spectral imaging for remote sensing, *Lincoln laboratory journal* 14 (2003) 3–28.
- P. Ghamisi, N. Yokoya, J. Li, W. Liao, S. Liu, J. Plaza, B. Rasti, A. Plaza, Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art, *IEEE Geoscience and Remote Sensing Magazine* 5 (2017) 37–78.
- L. M. Dale, A. Thewis, C. Boudry, I. Rotar, P. Dardenne, V. Baeten, J. A. F. Pierna, Hyperspectral imaging applications in agriculture and agro-food product quality and safety control: A review, *Applied Spectroscopy Reviews* 48 (2013) 142–159.
- J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, J. Chanussot, Hyperspectral remote sensing data analysis and future challenges, *IEEE Geoscience and Remote Sensing Magazine* 1 (2013) 6–36.
- T. Vo-Dinh, B. Cullum, P. Kasili, Development of a multi-spectral imaging system for medical applications, *Journal of Physics D: Applied Physics* 36 (2003) 1663.
- N. Yokoya, C. Grohnfeldt, J. Chanussot, Hyperspectral and multispectral data fusion: A comparative review of the recent literature, *IEEE Geoscience and Remote Sensing Magazine* 5 (2017) 29–56.
- J. Liu, Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details, *International Journal of Remote Sensing* 21 (2000) 3461–3472.
- B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, M. Selva, Mtf-tailored multiscale fusion of high-resolution ms and pan imagery, *Photogrammetric Engineering & Remote Sensing* 72 (2006) 591–596.
- B. Aiazzi, S. Baronti, M. Selva, Improving component substitution pansharpening through multivariate regression of ms + pan data, *IEEE Transactions on Geoscience and Remote Sensing* 45 (2007) 3230–3239.
- N. Yokoya, T. Yairi, A. Iwasaki, Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion, *IEEE Transactions on Geoscience and Remote Sensing* 50 (2012) 528–537.
- Q. Wei, N. Dobigeon, J. Tourneret, Fast fusion of multi-band images based on solving a Sylvester equation, *IEEE Transactions on Image Processing* 24 (2015) 4109–4121.
- J. Xie, Y. Wang, J. Li, Hyperspectral and multispectral data fusion with 1d-convolution on spectrum, in: *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 2022, pp. 2394–2397.
- F. Palsson, J. R. Sveinsson, M. O. Ulfarsson, Multispectral and hyperspectral image fusion using a 3-d-convolutional neural network, *IEEE Geoscience and Remote Sensing Letters* 14 (2017) 639–643.
- W.-S. Lai, J.-B. Huang, N. Ahuja, M.-H. Yang, Deep laplacian pyramid networks for fast and accurate super-resolution, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 624–632.
- X. Wang, X. Wang, K. Zhao, X. Zhao, C. Song, Fsl-unet: Full-scale linked unet with spatial-spectral joint perceptual attention for hyperspectral and multispectral image fusion, *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022) 1–14.
- Q. Xie, M. Zhou, Q. Zhao, D. Meng, W. Zuo, Z. Xu, Multispectral and hyperspectral image fusion by ms/hs fusion net, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1585–1594.
- Q. Xie, M. Zhou, Q. Zhao, Z. Xu, D. Meng, Mhf-net: An interpretable deep network for multispectral and hyperspectral image fusion, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020) 1–1.
- J. M. Ramirez, J. I. Martínez-Torre, H. Arguello, Ladmm-net: An unrolled deep network for spectral image fusion from compressive data, *Signal Processing* 189 (2021) 108239.
- R. Jacome, J. Bacca, H. Arguello, Deep-fusion: An end-to-end approach for compressive spectral image fusion, in: *2021 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2021, pp. 2903–2907.
- R. Dian, S. Li, X. Kang, Regularizing hyperspectral and multispectral image fusion by cnn denoiser, *IEEE Transactions on Neural Networks and Learning Systems* (2020) 1–12.
- X. Zhang, W. Huang, Q. Wang, X. Li, Ssr-net: Spatial-spectral reconstruction network for hyperspectral and multispectral image fusion, *IEEE Transactions on Geoscience and Remote Sensing* (2020) 1–13.
- B. Qiao, B. Xu, Y. Xie, Y. Lin, Y. Liu, X. Zuo, A. Loddo, Hmft: Hyperspectral and multispectral image fusion super-resolution method based on efficient transformer and spatial-spectral attention mechanism, *Intell. Neuroscience* 2023 (2023).
- S. Jia, Z. Min, X. Fu, Multiscale spatial-spectral transformer network for hyperspectral and multispectral image fusion, *Information Fusion* 96 (2023) 117–129.
- X. Wang, X. Wang, R. Song, X. Zhao, K. Zhao, Mct-net: Multi-hierarchical cross transformer for hyperspectral and multispectral image fusion, *Knowledge-Based Systems* 264 (2023) 110362.
- D. Ulyanov, A. Vedaldi, V. Lempitsky, Deep image prior, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9446–9454.
- O. Sidorov, J. Y. Hardeberg, Deep hyperspectral prior: denoising, inpainting, super-resolution, *arXiv preprint arXiv:1902.00301* (2019).
- J. Bacca, Y. Fonseca, H. Arguello, Compressive spectral image reconstruction using deep prior and low-rank tensor representation, *Applied optics* 60 (2021) 4197–4207.
- T. Gelvez, J. Bacca, H. Arguello, Interpretable deep image prior method inspired in linear mixture model for compressed spectral image recovery, in: *2021 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2021, pp. 1934–1938.
- X. Wang, R. A. Borsoi, C. Richard, J. Chen, Deep hyperspectral and multispectral image fusion with inter-image variability, *IEEE Transactions on Geoscience and Remote Sensing* (2023).
- Y. Zheng, J. Li, Y. Li, J. Guo, X. Wu, J. Chanussot, Hyperspectral pansharpening using deep prior and dual attention residual network, *IEEE transactions on geoscience and remote sensing* 58 (2020) 8059–8076.
- Z. Liu, Y. Zheng, X.-H. Han, Unsupervised multispectral and hyperspectral image fusion with deep spatial and spectral priors, in: *Proceedings of the Asian Conference on Computer Vision*, 2020.
- Z. Wang, B. Chen, R. Lu, H. Zhang, H. Liu, P. K. Varshney, Fusionnet: An unsupervised convolutional variational network for hyperspectral and multispectral image fusion, *IEEE Transactions on Image Processing* 29 (2020) 7565–7577.
- Y. Sun, J. Liu, J. Yang, Z. Xiao, Z. Wu, A deep image prior-based interpretable network for hyperspectral image fusion, *Remote Sensing Letters* 12 (2021) 1250–1259.
- S. Li, R. Dian, H. Liu, Learning the external and internal priors for multispectral and hyperspectral image fusion, *Science China Information Sciences* 66 (2023).
- G. Vivone, Multispectral and hyperspectral image fusion in remote sensing: A survey, *Information Fusion* 89 (2023) 405–417.
- J. Gu, H. Lu, W. Zuo, C. Dong, Blind super-resolution with iterative kernel correction, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1604–1613.
- Z. Liu, Y. Zheng, X.-H. Han, Deep unsupervised fusion learning for hyperspectral image super resolution, *Sensors* 21 (2021) 2348.

- J. Nie, L. Zhang, W. Wei, Z. Lang, Y. Zhang, Unsupervised alternating optimization for blind hyperspectral imagery super-resolution, arXiv preprint arXiv:2012.01745 (2020).
- W. G. C. Bandara, J. M. J. Valanarasu, V. M. Patel, Hyperspectral pansharpening based on improved deep image prior and residual reconstruction, *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021) 1–16.
- H. Arguello, J. Bacca, H. Kariyawasam, E. Vargas, M. Marquez, R. Hettiarachchi, H. Garcia, K. Herath, U. Haputhanthri, B. S. Ahluwalia, et al., Deep optical coding design in computational imaging: a data-driven framework, *IEEE Signal Processing Magazine* 40 (2023) 75–88.
- J. Gao, J. Li, M. Jiang, Hyperspectral and multispectral image fusion by deep neural network in a self-supervised manner, *Remote Sensing* 13 (2021) 3226.
- D. Hong, N. Yokoya, J. Chanussot, X. X. Zhu, An augmented linear mixing model to address spectral variability for hyperspectral unmixing, *IEEE Transactions on Image Processing* 28 (2018) 1923–1938.
- B. Rasti, P. Ghamisi, J. Plaza, A. Plaza, Fusion of hyperspectral and lidar data using sparse and low-rank component analysis, *IEEE Transactions on Geoscience and Remote Sensing* 55 (2017) 6354–6365.
- Q. Wei, N. Dobigeon, J.-Y. Tourneret, Fuse: A fast multi-band image fusion algorithm, in: 2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), IEEE, 2015, pp. 161–164.
- J. Bacca, H. Vargas, H. Arguello, A constrained formulation for compressive spectral image reconstruction using linear mixture models, in: Proc. Conf. CAMSAP, IEEE, 2017, pp. 1–5.
- J. M. Ramirez, H. Arguello, Multiresolution compressive feature fusion for spectral image classification, *IEEE Transactions on Geoscience and Remote Sensing* 57 (2019) 9900–9911.
- N. Akhtar, F. Shafait, A. Mian, Sparse spatio-spectral representation for hyperspectral image super-resolution, in: European conference on computer vision, Springer, 2014, pp. 63–78.
- M. Simoes, J. Bioucas-Dias, L. B. Almeida, J. Chanussot, A convex formulation for hyperspectral image superresolution via subspace-based regularization, *IEEE Transactions on Geoscience and Remote Sensing* 53 (2014) 3373–3388.
- J. Bacca, C. V. Correa, H. Arguello, Noniterative hyperspectral image reconstruction from compressive fused measurements, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12 (2019) 1231–1239.
- Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity, *IEEE Transactions on Image Processing* 13 (2004) 600–612.
- M. V. Narkhede, P. P. Bartakke, M. S. Sutaone, A review on weight initialization strategies for neural networks, *Artificial intelligence review* 55 (2022) 291–322.
- D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
- Jet Propulsion Laboratory, NASA, 2006-2019 AVIRIS Data Portal, <https://aviris.jpl.nasa.gov/dataportal/>, 2019. [Online; accessed 19-February-2020].
- G. de Inteligencia Computacional, Hyper Remote Sensing Scenes, <http://www.ehu.eus/>, 2008.
- S. Shi, L. Zhang, J. Chen, Hyperspectral and multispectral image fusion using the conditional denoising diffusion probabilistic model, arXiv preprint arXiv:2307.03423 (2023).
- J.-F. Hu, T.-Z. Huang, L.-J. Deng, T.-X. Jiang, G. Vivone, J. Chanussot, Hyperspectral image super-resolution via deep spatio-spectral attention convolutional neural networks, *IEEE Transactions on Neural Networks and Learning Systems* 33 (2021) 7251–7265.
- H. Arguello, S. Pinilla, Y. Peng, H. Ikoma, J. Bacca, G. Wetzstein, Shift-variant color-coded diffractive spectral imaging system, *Optica* 8 (2021) 1424–1434.
- J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, J. Paisley, Pannet: A deep network architecture for pan-sharpening, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 5449–5457.
- X. Zhang, W. Huang, Q. Wang, X. Li, Ssr-net: Spatial-spectral reconstruction network for hyperspectral and multispectral image fusion, *IEEE Transactions on Geoscience and Remote Sensing* 59 (2020) 5953–5965.
- W. Wang, X. Fu, W. Zeng, L. Sun, R. Zhan, Y. Huang, X. Ding, Enhanced deep blind hyperspectral image fusion, *IEEE transactions on neural networks and learning systems* (2021).
- T. Gelvez-Barrera, J. Bacca, H. Arguello, Mixture-net: low-rank deep image prior inspired by mixture models for spectral image recovery, *Signal Processing* (2023) 109296.