

# Deep Hyperspectral and Multispectral Image Fusion with Inter-image Variability

Xiuheng Wang, *Student Member, IEEE*, Ricardo Augusto Borsoi, *Member, IEEE*,  
Cédric Richard, *Senior Member, IEEE*, and Jie Chen, *Senior Member, IEEE*.

**Abstract**—Hyperspectral and multispectral image fusion allows us to overcome the hardware limitations of hyperspectral imaging systems inherent to their lower spatial resolution. Nevertheless, existing algorithms usually fail to consider realistic image acquisition conditions. This paper presents a general imaging model that considers inter-image variability of data from heterogeneous sources and flexible image priors. The fusion problem is stated as an optimization problem in the maximum a posteriori framework. We introduce an original image fusion method that, on the one hand, solves the optimization problem accounting for inter-image variability with an iteratively reweighted scheme and, on the other hand, that leverages lightweight CNN-based networks to learn realistic image priors from data. In addition, we propose a zero-shot strategy to directly learn the image-specific prior of the latent images in an unsupervised manner. The performance of the algorithm is illustrated with real data subject to inter-image variability.

**Index Terms**—Hyperspectral data, multispectral data, inter-image variability, image fusion, deep learning, zero-shot.

## I. INTRODUCTION

Hyperspectral imaging systems acquire scenes by recording hundreds of narrow, contiguous spectral bands ranging from visible up to infrared wavelengths. Their rich spectral information has attracted interest in many applications such as remote sensing for mineral exploration, vegetation monitoring, and land cover analysis [1]. Nevertheless, the high spectral resolution of hyperspectral images (HIs) limits their spatial resolution because of hardware limitations [2]. In contrast, multispectral cameras can achieve a much higher spatial resolution but over a small number of spectral bands. Consequently, a strategy to improve the spatial resolution of HIs is to fuse them with multispectral images (MIs) of the same scene. This results in the hyperspectral and multispectral image fusion (HMIF) problem.

Several strategies have been proposed to solve the HMIF problem. These strategies can be roughly divided into component substitution or multiresolution analysis methods, matrix or tensor factorization methods, and deep learning approaches. Component substitution or multiresolution analysis methods aim to substitute some patterns of the HI, high-frequency ones in particular, by information extracted from the MI [3], [4], [5]. These techniques employ different representations of the

images, e.g., in the wavelet domain, which are also used for pansharpening [6], [7].

Subspace-based formulations have become very popular to address HMIF problems since they significantly reduce their dimensionality [3], [8]. They also have a close connection with the widely used linear mixing model [9], [10], which represents each pixel of an HI as a linear combination of a small number of spectral signatures. Several subspace-based formulations have been proposed, often employing prior information about the basis vectors or their contributions in the decomposition, to improve the results. Examples include sparse dictionary learning [11], [12] or matrix factorization [3] approaches, which can use, e.g., spatial [8] and sparse [13], [14] regularizers or patch-level processing [15]. Efficient algorithms also convert this problem into solving a Sylvester equation [16]. Some approaches have considered the manifold structure of the image patches [17]. Other approaches have explored the representation of HIs and MIs as three dimensional tensors [18], [19], [20]. Low-rank tensor models have been used to represent the high-resolution images (HRIs), such as the canonical polyadic decomposition [18], the Tucker decomposition [19], [20], [21], and the block term decomposition [22].

Deep learning approaches have recently become very popular for HMIF [23], [24], [25]. These approaches leverage the capability of neural networks to represent complex signals and images. Early supervised approaches were based upon classical neural network architectures used in image processing such as 3D convolutional neural networks (CNN) [26], while more recent methods explore physical acquisition models to design architectures with improved interpretability [27], e.g., incorporating CNN results as priors in model-based frameworks [28], [29] or using architectures inspired by unrolling principle [30]. However, the scarcity of training data with ground truth has motivated the development of unsupervised approaches, that depend only on the observed HI and MI. Examples include the use of autoencoders with shared weights [31], [32], [33], and approaches based on deep image priors [34], which parameterize the HRI as the output of a neural network and train the latter using different options for the network inputs [35], [36].

Although different strategies have been investigated to solve the HMIF problem, these methods assume that the observed HI and MI are acquired at the same time instant and under the same conditions. However, platforms carrying both hyperspectral and multispectral imaging systems are still limited [37]. On the contrary, due to the wider availability of satellites

Xiuheng Wang and Cédric Richard are with Université Côte d’Azur, CNRS, OCA, F-06108, Nice, France (e-mail: xiuheng.wang@oca.eu, cedric.richard@unice.fr). Ricardo Augusto Borsoi is with Université de Lorraine, CNRS, CRAN, F-54000, Nancy, France (e-mail: raborsoi@gmail.com). Jie Chen is with Centre of Intelligent Acoustics and Immersive Communications at School of Marine Science and Technology, Northwestern Polytechnical University, Xi’an, 710072, China (e-mail: dr.jie.chen@ieec.org).

with multispectral sensors, e.g., the Sentinel, Landsat and Quickbird missions, it has become of great interest to fuse HIs and MIs acquired at different time instants by different instruments [38]. When applied in these realistic conditions, most existing methods suffer from severe limitations as they ignore variability between the HI and MI. Inter-image variability includes localized spatial and spectral changes and can occur due to differences in acquisition conditions caused by, e.g., atmospheric, illumination or seasonal variations [39], as well as abrupt changes [40].

To tackle this issue, several HMIF frameworks addressing inter-image variability have been recently proposed [37], [21], [41], [42], [43], [44], [45]. A detailed review of these methods is provided in Section II. These methods formulate the HMIF problem with a key difference when compared to the original approaches: the HI and the MI are assumed to be generated from distinct HRIs, which are allowed to be different because of spatially homogeneous variations [37], [41] or spatially localized ones [21]. However, considering inter-image variability renders the HMIF problem significantly more ill-posed, which makes the use of appropriate prior information about the HRIs very important in order to achieve good performance.

Existing HMIF works that consider inter-image variability rely on handcrafted priors, such as low-rank matrix [37] or tensor [21], [41] decompositions. However, these priors are not adequate to model complex contents embedded in real HIs. Without considering inter-image variability, this issue has been addressed in the HMIF problem by exploring the powerful representation capability of deep learning methods, as noted by various recent works on this topic. Nevertheless, devising learning-based approaches to address inter-image variability in HMIF incurs additional challenges, first because very little data is available for training. Indeed, since inter-image variability originates from complex physical phenomena, it is difficult to generate realistic synthetic data to be used for training even if HIs of a single scene are available. This makes learning an end-to-end mapping from an HI and an MI to the HRIs unfeasible.

Recently, deep image priors [34] and plug-and-play strategies [46] have been used to introduce prior information with either pre-trained or unsupervised neural networks. However, adequately addressing inter-image variability requires considering two different HRIs, underlying the HI and the MI, respectively. Thus, directly exploiting such strategies to address inter-image variability in HMIF is not very effective since: 1) existing strategies in this category would fail to account for the joint prior information between the two HRIs, and 2) each of the images can have distinct statistical properties, which makes obtaining adequate priors more difficult. Moreover, although deep image priors are unsupervised [34], they require careful setup of the network architecture and the number of stochastic gradient iterations to produce reasonable results. It must be noted that these challenges related to the lack of training data and the corresponding difficulty in learning priors of the scene of interest are also encountered more generally in HMIF, i.e., even when inter-image variability is not present.

In this paper, we propose a new image fusion method accounting for inter-image variability between HIs and MIs

which addresses the aforementioned challenges. First, to adequately represent the image-specific information as well as the joint prior information between the two HRIs, we propose a mixture distribution that accounts for the leptokurtic nature of the inter-image variations while, at the same time, represents complex image content by implicitly exploiting learning-based image priors. An iteratively reweighted optimization strategy is then proposed, and the regularization by denoising (RED) [47] framework is employed to implicitly introduce prior information about the HRIs by means of denoising engines, one for each latent HRI. The denoisers are trained using a zero-shot strategy [48] and adapted during the optimization process, which allows them to account for the content of each individual HRI. The proposed algorithm is called *Deep hyperspectral and multispectral Image Fusion with Inter-image Variability* (DIFIV). Experiments on data with real inter-image variability demonstrate the superiority of DIFIV compared to other state-of-the-art methods. The contributions of the paper are summarized as follows.

- A general imaging model is formulated, where the inter-image variations of the HRIs are modeled by a hyper-Laplacian distribution to account for the joint image content, while the image content specific to each HRI is learned by two distinct deep CNNs.
- To solve the non-convex, non-smooth HMIF optimization problem, a variable splitting strategy is combined with an iteratively reweighted scheme to tackle the difficulties introduced by both the hyper-Laplacian and deep priors, which are defined implicitly based on CNN denoisers under the RED framework.
- We use a zero-shot strategy inspired by [48] to learn the CNN denoisers based only on the observed HI and MI. Moreover, unlike the original use of zero-shot methods for single image restoration, the denoisers are trained iteratively during the optimization process based on the currently estimated HRIs. This allows the learned priors to represent the individual information in each of the HRIs adaptively while incorporating at the same time information from both low resolution images as the method converges. Furthermore, the architecture of CNNs is made lightweight by considering separable convolutions and a low-rank representation of HIs to yield a small number of network parameters.

The paper is organized as follows. In Section II, the HI and MI observation processes are presented, as well as a review of recent methods considering inter-image variability. Section III formulates a new model and introduces the proposed method. Experimental results with data containing real inter-image variability are given in Section IV. Finally, Section V concludes the paper.

## II. IMAGE FUSION WITH INTER-IMAGE VARIABILITY

Let us denote an HI with  $L_h$  bands and  $N$  pixels by  $\mathbf{Y}_h \in \mathbb{R}^{L_h \times N}$ , and a MI with  $L_m$  bands and  $M$  pixels by  $\mathbf{Y}_m \in \mathbb{R}^{L_m \times M}$ , where  $L_m < L_h$  and  $N < M$ . These images are assumed to be degraded versions of a pair of underlying HRIs  $\mathbf{Z}_h \in \mathbb{R}^{L_h \times M}$  and  $\mathbf{Z}_m \in \mathbb{R}^{L_h \times M}$  with high spatial

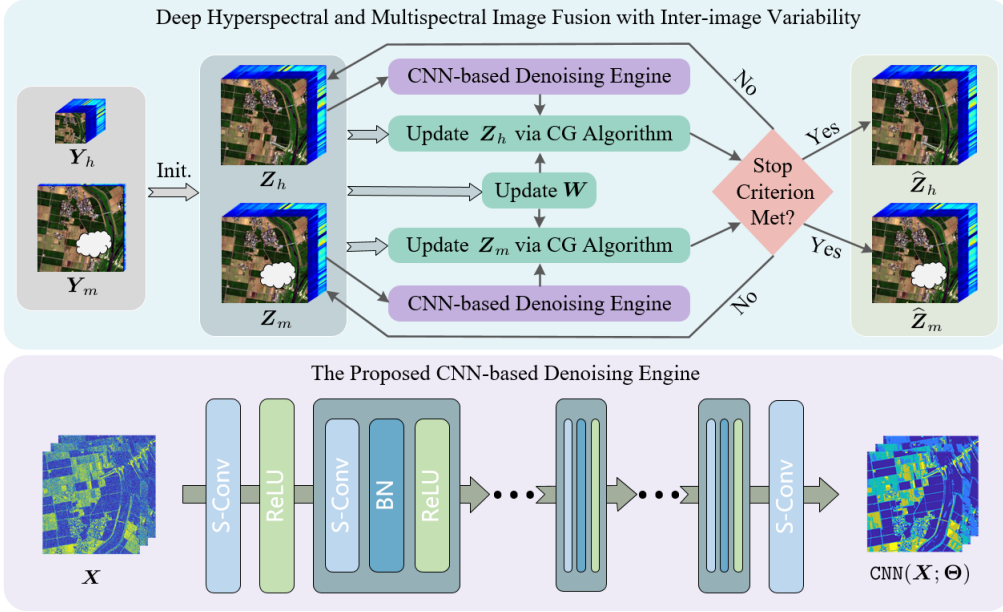


Figure 1. **Top panel:** Overall illustration of the proposed Deep hyperspectral and multispectral Image Fusion with Inter-image Variability (DIFIV) method: the HRIs underlying the HI and MI ( $Z_h$  and  $Z_m$ ) are initialized (Init.), with interpolations of observed images ( $Y_h$  and  $Y_m$ ), and then used to compute the inter-image variability weighting term  $W$  and to update the CNN-based denoisers; afterwards, these are used to re-compute the HRIs using a conjugate-gradient based algorithm; this process is repeated iteratively until convergence. **Bottom panel:** The neural network architecture of our CNN-based denoising engine (S-Conv, BN, and ReLU stand for separable convolution, batch normalization and rectified linear unit layers, respectively).

and spectral resolutions, which are related according to the following model:

$$\begin{aligned} Y_h &= Z_h F D + E_h, \\ Y_m &= R Z_m + E_m, \end{aligned} \quad (1)$$

in which matrices  $F \in \mathbb{R}^{M \times M}$  and  $D \in \mathbb{R}^{M \times N}$  represent optical blurring and spatial downsampling occurring at the hyperspectral sensor, respectively; matrix  $R \in \mathbb{R}^{L_m \times L_h}$  contains the spectral response functions (SRF) of the multispectral instrument, and  $E_h \in \mathbb{R}^{L_h \times N}$  and  $E_m \in \mathbb{R}^{L_m \times M}$  denote additive noise.

In this setting, the image fusion problem consists of recovering the HRIs  $Z_h$  and  $Z_m$  given the observations  $Y_h$  and  $Y_m$ . Most of the previous methods consider that  $Y_h$  and  $Y_m$  are degraded from the same source, i.e.,  $Z_h = Z_m$ , which intrinsically assumes that they are acquired under the same conditions, e.g., by sensors on board a single satellite. However, due to the wider availability of satellites equipped with multispectral sensors, it is of great interest to fuse HIs and MIs acquired by different instruments at different time instants [38]. In that case, by assuming that  $Z_h = Z_m$ , most existing methods ignore variabilities between the HI and MI, which can occur due to differences in acquisition conditions caused by, e.g., atmospheric, illumination or seasonal variations [39], or abrupt changes [40].

Recently, image fusion frameworks addressing inter-image variability have been proposed in [37], [21], [41], [42], [43], [44], [45]. Such methods estimate both HRIs  $Z_h$  and  $Z_m$  by using different assumptions to model both the images and the inter-image changes. The first method to address this problem was FuVar [37]. It considers that the HRIs satisfy the linear

mixing model (LMM) [9], but with a distinct set of spectral basis vectors for each image:

$$Z_h = M_h A, \quad Z_m = M_m A, \quad (2)$$

where  $M_h$  and  $M_m \in \mathbb{R}^{L_h \times R}$  denote the set of spectral basis vectors related to the HI and MI, respectively, and  $A \in \mathbb{R}^{R \times N}$  their corresponding spatial coefficients. Note that  $M_h$  and  $M_m$  are associated with the spectral signatures of the pure materials (i.e., the endmembers) in the HI and MI, respectively. FuVar considers  $M_h$  and  $M_m$  to be related to one another through a set of smooth multiplicative scaling factors  $\Phi \in \mathbb{R}^{L_h \times R}$  [49]:

$$M_m = M_h \odot \Phi, \quad (3)$$

where  $\odot$  denotes the Hadamard product. Thus, this model successfully accounts for changes in the spectral signatures of the endmembers between the HI and the MI, which can occur when the materials are affected by seasonal variations or when the MI is affected by uniform changes caused by, e.g., different illumination conditions. However, the coefficients  $A$  shared by both images limit the capability of FuVar to represent inter-image changes in the spatial domain.

This limitation has been addressed by considering spatially and spectrally localized inter-image variations through an additive model in a tensor-based framework [21]. This latter work considers a model of the form:

$$Z_h, \quad Z_m = Z_h + \Psi, \quad (4)$$

where  $\Psi \in \mathbb{R}^{L_h \times M}$  denotes a set of additive variability factors. Both the HRI  $Z_h$  and the variability  $\Psi$  are assumed to admit a Tucker tensor decomposition with low multilinear ranks [50]. This reduces the dimensionality of the problem

and allows theoretical identifiability and recovery guarantees to be obtained [21].

A related work proposes to jointly address the image fusion and hyperspectral unmixing problems in the presence of inter-image spectral variability [41]. This consists of the recovery of both the HRIs and the spectral signatures of the endmembers and their abundances. An LL1 tensor model is considered, which is closely related to the LMM in (2) but involves an additional low-rank assumption on the coefficient maps  $\mathbf{A}$  that allows theoretical identifiability results to be derived. Other works propose to consider intra-image variability by extending the LMM to consider spatial endmember variability, i.e., variability within a single image [43], [44]. Another work considers a robust version of the data fidelity term related to the MI in the cost function to reduce the impact of possible changes or outliers in the image fusion process [45]. However, these methods still assume that the HRIs underlying the HI and the MI are equal.

Despite the success of these approaches in addressing the inter-image variability problem, they all rely on handcrafted priors for the HR images  $\mathbf{Z}_h$  and  $\mathbf{Z}_m$ , which limits their capability of representing realistic image content. In this work, we propose an image fusion method that leverages the expressive power of CNNs in order to construct accurate image priors for the HRIs while accounting for inter-image variability, as detailed in the following section.

### III. THE PROPOSED METHOD

The proposed image fusion method is based on three important axes/contributions: 1) an imaging model that incorporates inter-image variability with learned image priors, 2) an optimization scheme that can handle these flexible penalties, 3) a lightweight unsupervised (zero-shot) scheme to iteratively learn deep priors of the latent HRIs during the reconstruction process. The proposed image fusion method is presented through four steps. First, we present the imaging model in Subsection III-A and formulate the optimization problem. In Subsection III-B we describe an iteratively reweighted scheme to optimize the cost function. The optimization steps, as well as the integration of deep priors, are described in Subsection III-C. We then address the design of CNN architecture and its image-adapted training strategy in Subsection III-D. An overall illustration of the proposed DIFIV method is shown in Figure 1.

#### A. The imaging model

Using a probabilistic framework, the HMIF problem can be formulated as the recovery of the mean or mode of the posterior probability distribution function (PDF)  $p(\mathbf{Z}_h, \mathbf{Z}_m | \mathbf{Y}_h, \mathbf{Y}_m)$  of both HRIs given the LR observations. Using Bayes theorem, this PDF can be written as:

$$p(\mathbf{Z}_h, \mathbf{Z}_m | \mathbf{Y}_h, \mathbf{Y}_m) \propto p(\mathbf{Y}_h | \mathbf{Z}_h) p(\mathbf{Y}_m | \mathbf{Z}_m) p(\mathbf{Z}_m, \mathbf{Z}_h), \quad (5)$$

where we assumed the HI and MI to be conditionally independent given their high-resolution counterparts.

The likelihoods of the observed images  $\mathbf{Y}_h$  and  $\mathbf{Y}_m$  can be written according to their data generation process in (1). More precisely, assuming the elements of  $\mathbf{E}_h$  and  $\mathbf{E}_m$  to be i.i.d. Gaussian random variables with variance  $\sigma_h^2$  and  $\sigma_m^2$ , respectively, the conditional distributions of  $\mathbf{Y}_m$  and  $\mathbf{Y}_n$  in (5) are given by:

$$p(\mathbf{Y}_h | \mathbf{Z}_h) = \mathcal{MN}(\mathbf{Z}_h \mathbf{F} \mathbf{D}, \sigma_h^2 \mathbf{I}_{L_h}, \mathbf{I}_N), \quad (6)$$

$$p(\mathbf{Y}_m | \mathbf{Z}_m) = \mathcal{MN}(\mathbf{R} \mathbf{Z}_m, \sigma_m^2 \mathbf{I}_{L_m}, \mathbf{I}_M), \quad (7)$$

where  $\mathcal{MN}(\mathbf{Y}, \mathbf{\Sigma}_r, \mathbf{\Sigma}_c)$  denotes the matrix normal distribution with mean matrix  $\mathbf{Y}$  and row and column covariance matrices  $\mathbf{\Sigma}_r$  and  $\mathbf{\Sigma}_c$ , respectively [16].

The challenging question concerns how to meaningfully define the joint prior  $p(\mathbf{Z}_m, \mathbf{Z}_h)$  for both HRIs. This question is not trivial when the images differ due to acquisition conditions or seasonal variations. A simplistic possibility is to consider the images to be independent and to use priors used for super-resolution without variability, such as low-rank matrix and tensor models [3], [18], [51], piecewise-smoothness [8] or learned deep priors [29], [28], [52]. However, the images  $\mathbf{Z}_m$  and  $\mathbf{Z}_h$  are observations of the same scene, and thus are strongly dependent. Considering this, we can state the following desirable properties for  $p(\mathbf{Z}_m, \mathbf{Z}_h)$ :

- Apart from possible smooth inter-image variations (such as, e.g., illumination or atmospheric changes, which tend to impact the images uniformly [39]), changes between  $\mathbf{Z}_m$  and  $\mathbf{Z}_h$  are generally small and sparse; high magnitude changes are concentrated in a relatively small number of pixels and bands [40].
- The prior should promote images  $\mathbf{Z}_m$  and  $\mathbf{Z}_h$  which are statistically similar to real hyperspectral images (e.g., they can be well represented by learned priors).

To achieve the above desiderata, we consider a mixture distribution, given by:

$$\log p(\mathbf{Z}_m, \mathbf{Z}_h) \propto -\frac{\lambda_p}{2} \sum_{\ell, n} |\delta_h^{(\ell, n)} - \delta_m^{(\ell, n)}|^p - \lambda_m \phi_m(\mathbf{Z}_m) - \lambda_h \phi_h(\mathbf{Z}_h), \quad (8)$$

for  $0 < p \leq 1$ , where  $\delta_h^{(\ell, n)}$  and  $\delta_m^{(\ell, n)}$  denote the  $(\ell, n)$ -th locations of a high-pass spatio-spectral filtered version of  $\mathbf{Z}_h$  and  $\mathbf{Z}_m$ , which are denoted by  $\mathbf{\Delta}_h$  and  $\mathbf{\Delta}_m$ , respectively. We assume this filtering to be computed through an operator  $\mathcal{G}$  satisfying  $\mathbf{\Delta}_h = \mathcal{G}(\mathbf{Z}_h)$ ,  $\mathbf{\Delta}_m = \mathcal{G}(\mathbf{Z}_m)$ , and in vector form as  $\text{vec}(\mathbf{\Delta}_h) = \mathbf{G} \text{vec}(\mathbf{Z}_h)$  and  $\text{vec}(\mathbf{\Delta}_m) = \mathbf{G} \text{vec}(\mathbf{Z}_m)$  where  $\mathbf{G}$  is the matrix form of  $\mathcal{G}$ . One natural example for  $\mathcal{G}$  is the spatio-spectral gradient operator, e.g. Laplacian filter. Parameters  $\lambda_p$ ,  $\lambda_m$  and  $\lambda_h$  are regularization parameters.

The first term in (8) corresponds to an i.i.d. hyper-Laplacian distribution for the difference between the filtered HRIs [53], which has also been previously used to represent the gradient of the HRI in image fusion [54]. This distribution is effective for modeling leptokurtic (i.e., heavy-tailed) distributions such as images [53]. This can represent an important characteristic of the inter-image changes since these can be restricted to a comparatively small number of pixels and are concentrated at low-frequency spatial content [55]. The functions  $\phi_h(\cdot)$  and



$\phi_m(\cdot)$  encode prior knowledge about each HRI, and will be learned implicitly by using deep CNNs.

Note that the prior in (8) also corresponds to a model for the inter-image variability, which can be written as:

$$\mathbf{Z}_m = \mathbf{Z}_h + \Psi_\Delta. \quad (9)$$

What is distinctive in (9) when compared to the model in (4) is how prior information is chosen. The prior for the inter-image variability term  $\Psi_\Delta$  cannot be written in an analytical form; instead, its properties follow from the interactions of the different terms in (8). The first term encourages the inter-image variability  $\Psi_\Delta$  to have small and sparse gradients. The last two terms employ CNNs that can incorporate realistic prior information about each of the HRIs, and only constrain  $\Psi_\Delta$  indirectly through its effect on  $\mathbf{Z}_m$  and  $\mathbf{Z}_h$ .

Given this model, the image fusion problem then consists of finding the HRIs  $\mathbf{Z}_h$  and  $\mathbf{Z}_m$  which maximize the logarithm of the posterior distribution  $p(\mathbf{Z}_h, \mathbf{Z}_m | \mathbf{Y}_h, \mathbf{Y}_m)$  defined in (5). This corresponds to the following optimization problem:

$$\begin{aligned} \min_{\mathbf{Z}_h, \mathbf{Z}_m} & \frac{1}{2} \|\mathbf{Y}_h - \mathbf{Z}_h \mathbf{F} \mathbf{D}\|_F^2 + \frac{1}{2} \|\mathbf{Y}_m - \mathbf{R} \mathbf{Z}_m\|_F^2 \\ & + \lambda_h \phi_h(\mathbf{Z}_h) + \lambda_m \phi_m(\mathbf{Z}_m) \\ & + \frac{\lambda_p}{2} \|\mathcal{G}(\mathbf{Z}_h) - \mathcal{G}(\mathbf{Z}_m)\|_{p,p}^p, \end{aligned} \quad (10)$$

where  $\|\cdot\|_{p,p}$  is the entrywise  $L_p$  matrix norm, satisfying  $\|\mathcal{G}(\mathbf{Z}_h) - \mathcal{G}(\mathbf{Z}_m)\|_{p,p}^p = \sum_{\ell,n} |\delta_h^{(\ell,n)} - \delta_m^{(\ell,n)}|^p$ . The spatial and spectral priors of  $\mathbf{Z}_m$  and  $\mathbf{Z}_h$  are encoded in  $\phi_h(\mathbf{Z}_h)$  and  $\phi_m(\mathbf{Z}_m)$ , respectively.

### B. An iteratively reweighted update scheme

Optimizing the cost function in (10) is challenging. Apart from the image priors  $\phi_h(\cdot)$  and  $\phi_m(\cdot)$  that will be defined in the sequel, the inter-image prior term (i.e., the last term in (10)) is, in general, a non-convex and non-smooth function of both  $\mathbf{Z}_h$  and  $\mathbf{Z}_m$ , which is not straightforward to optimize. To address this problem, we consider an iteratively reweighted optimization strategy [56], [57]. First, note that the last term in (10) can be written as:

$$\sum_{\ell,n} |\delta_h^{(\ell,n)} - \delta_m^{(\ell,n)}|^p = \sum_{\ell,n} w_{\ell,n} |\delta_h^{(\ell,n)} - \delta_m^{(\ell,n)}|^2, \quad (11)$$

where the weights  $w_{\ell,n}$  are given by

$$w_{\ell,n} = |\delta_h^{(\ell,n)} - \delta_m^{(\ell,n)}|^{p-2}. \quad (12)$$

Since  $w_{\ell,n} \geq 0$ , (11) can be expressed as:

$$\sum_{\ell,n} w_{\ell,n} |\delta_h^{(\ell,n)} - \delta_m^{(\ell,n)}|^2 = \|\mathbf{W} \odot (\Delta_h - \Delta_m)\|_F^2, \quad (13)$$

where  $\mathbf{W}$  is a matrix whose  $(\ell, n)$ -th entry is given by  $\sqrt{w_{\ell,n}}$ , and  $\odot$  denotes the Hadamard product.

When matrix  $\mathbf{W}$  is fixed, (13) becomes a quadratic function of the HRIs, which can be effectively optimized. The nonlinear dependency of  $\mathbf{W}$  on  $\mathbf{Z}_h$  and  $\mathbf{Z}_m$  will be resolved by using an iterative strategy: first the cost function is optimized considering  $\mathbf{W}$  fixed to obtain  $\mathbf{Z}_h$  and  $\mathbf{Z}_m$ , and afterwards  $\mathbf{W}$  is updated according to an approximate version of (12)

by using the values of  $\mathbf{Z}_h$  and  $\mathbf{Z}_m$  computed from previous iteration [56]. This leads to the following iterative procedure, which is repeated until convergence:

1) For a fixed  $\mathbf{W}$ , compute  $\mathbf{Z}_h$  and  $\mathbf{Z}_m$  by solving the following optimization problem:

$$\begin{aligned} \min_{\mathbf{Z}_h, \mathbf{Z}_m, \Delta_h, \Delta_m} & \frac{1}{2} \|\mathbf{Y}_h - \mathbf{Z}_h \mathbf{F} \mathbf{D}\|_F^2 + \frac{1}{2} \|\mathbf{Y}_m - \mathbf{R} \mathbf{Z}_m\|_F^2 \\ & + \lambda_h \phi_h(\mathbf{Z}_h) + \lambda_m \phi_m(\mathbf{Z}_m) \\ & + \frac{\lambda_p}{2} \|\mathbf{W} \odot (\Delta_h - \Delta_m)\|_F^2 \\ \text{s.t. } & \Delta_h = \mathcal{G}(\mathbf{Z}_h), \Delta_m = \mathcal{G}(\mathbf{Z}_m). \end{aligned} \quad (14)$$

2) Update the entries of  $\mathbf{W}$  according to

$$w_{\ell,n} = (|\delta_h^{(\ell,n)} - \delta_m^{(\ell,n)}| + \epsilon)^{p-2}, \quad (15)$$

where  $\epsilon > 0$  is a small constant included in (12) to ensure the numerical stability of the algorithm.

3) Return to step 1) and repeat until convergence.

This strategy is efficient to solve sparsity-regularized optimization problems [58]. Moreover, iteratively reweighted optimization schemes have been shown to converge to a local stationary point under relatively mild conditions [56].

In the following subsection, we shall focus on the minimization problem (14).

### C. The optimization problem

Handcrafting powerful regularizers  $\phi_h(\mathbf{Z}_h)$  and  $\phi_m(\mathbf{Z}_m)$  along with solving the associated optimization problems efficiently is not a trivial task. In this subsection, we propose to learn the image prior directly from the observed data and incorporate it into the model-based optimization (14) to avoid designing regularizers analytically.

First, by introducing two auxiliary variables,  $\mathbf{V}_h = \mathbf{Z}_h$  and  $\mathbf{V}_m = \mathbf{Z}_m$ , problem (14) can be rewritten equivalently as:

$$\begin{aligned} \min_{\Omega} & \frac{1}{2} \|\mathbf{Y}_h - \mathbf{Z}_h \mathbf{F} \mathbf{D}\|_F^2 + \frac{1}{2} \|\mathbf{Y}_m - \mathbf{R} \mathbf{Z}_m\|_F^2 \\ & + \frac{\lambda_p}{2} \|\mathbf{W} \odot (\Delta_h - \Delta_m)\|_F^2 + \lambda_m \phi_m(\mathbf{V}_m) \\ & + \lambda_h \phi_h(\mathbf{V}_h) \\ \text{s.t. } & \mathbf{V}_h = \mathbf{Z}_h, \mathbf{V}_m = \mathbf{Z}_m, \\ & \Delta_h = \mathcal{G}(\mathbf{Z}_h), \Delta_m = \mathcal{G}(\mathbf{Z}_m), \end{aligned} \quad (16)$$

where  $\Omega = \{\mathbf{Z}_h, \mathbf{Z}_m, \mathbf{V}_h, \mathbf{V}_m, \Delta_h, \Delta_m\}$ . By using the half-quadratic splitting (HQS) approach [59], we can decouple the data fidelity and regularization terms in (16) and write this cost function as:

$$\begin{aligned} \mathcal{L}_\rho(\mathbf{Z}_h, \mathbf{Z}_m, \mathbf{V}_h, \mathbf{V}_m) & = \frac{1}{2} \|\mathbf{Y}_h - \mathbf{Z}_h \mathbf{F} \mathbf{D}\|_F^2 \\ & + \frac{1}{2} \|\mathbf{Y}_m - \mathbf{R} \mathbf{Z}_m\|_F^2 + \frac{\lambda_p}{2} \|\mathbf{W} \odot (\mathcal{G}(\mathbf{Z}_h) - \mathcal{G}(\mathbf{Z}_m))\|_F^2 \\ & + \frac{\rho}{2} \|\mathbf{Z}_m - \mathbf{V}_m\|_F^2 + \frac{\rho}{2} \|\mathbf{Z}_h - \mathbf{V}_h\|_F^2 \\ & + \lambda_m \phi_m(\mathbf{V}_m) + \lambda_h \phi_h(\mathbf{V}_h), \end{aligned} \quad (17)$$

with  $\rho \in \mathbb{R}_+$  the penalty parameter. In the following, we consider a block coordinate descent (BCD) strategy and minimize  $\mathcal{L}_\rho$  with respect to each variable, one at a time.

**Optimization w.r.t.  $\mathbf{Z}_h$ :** This optimization problem can be written as:

$$\min_{\mathbf{Z}_h} \frac{1}{2} \|\mathbf{Y}_h - \mathbf{Z}_h \mathbf{F} \mathbf{D}\|_F^2 + \frac{\lambda_p}{2} \|\mathbf{W} \odot (\mathcal{G}(\mathbf{Z}_h) - \mathcal{G}(\mathbf{Z}_m))\|_F^2 + \frac{\rho}{2} \|\mathbf{Z}_h - \mathbf{V}_h\|_F^2. \quad (18)$$

By taking the derivative of the cost function in (18), setting it equal to zero and using the vectorization property of matrix products, we obtain:

$$\begin{aligned} & - [(\mathbf{F} \mathbf{D})^\top \otimes \mathbf{I}_L]^\top \left( \text{vec}(\mathbf{Y}_h) - [(\mathbf{F} \mathbf{D})^\top \otimes \mathbf{I}_L] \text{vec}(\mathbf{Z}_h) \right) \\ & + \lambda_p \mathbf{G}^\top \text{diag}(\text{vec}(\mathbf{W}))^2 \mathbf{G} (\text{vec}(\mathbf{Z}_h - \mathbf{Z}_m)) \\ & + \rho \text{vec}(\mathbf{Z}_h - \mathbf{V}_h) = \mathbf{0}. \end{aligned} \quad (19)$$

Using the properties of the Kronecker product, this equation can be written as:

$$\begin{aligned} & \left( [(\mathbf{F} \mathbf{D})(\mathbf{F} \mathbf{D})^\top \otimes \mathbf{I}_L] + \lambda_p \mathbf{G}^\top \text{diag}(\text{vec}(\mathbf{W}))^2 \mathbf{G} \right. \\ & \left. + \rho \mathbf{I} \right) \text{vec}(\mathbf{Z}_h) = [(\mathbf{F} \mathbf{D})^\top \otimes \mathbf{I}_L]^\top \text{vec}(\mathbf{Y}_h) \\ & + \lambda_p \mathbf{G}^\top \text{diag}(\text{vec}(\mathbf{W}))^2 \mathbf{G} \text{vec}(\mathbf{Z}_m) + \rho \text{vec}(\mathbf{V}_h). \end{aligned} \quad (20)$$

which is a linear system of equations. However, solving this system directly is prohibitive due to its large dimension. Since the matrix on the left-hand side is symmetric positive-definite, we propose to solve this problem using the conjugate gradient (CG) algorithm, which requires only matrix-vector products that can be implemented implicitly and more efficiently.

**Optimization w.r.t.  $\mathbf{Z}_m$ :** This optimization problem can be written as:

$$\min_{\mathbf{Z}_m} \frac{1}{2} \|\mathbf{Y}_m - \mathbf{R} \mathbf{Z}_m\|_F^2 + \frac{\lambda_p}{2} \|\mathbf{W} \odot (\mathcal{G}(\mathbf{Z}_h) - \mathcal{G}(\mathbf{Z}_m))\|_F^2 + \frac{\rho}{2} \|\mathbf{V}_m - \mathbf{Z}_m\|_F^2. \quad (21)$$

Following the same steps as for problem (18), we obtain:

$$\begin{aligned} & \left( [\mathbf{I}_N \otimes \mathbf{R}^\top \mathbf{R}] + \lambda_p \mathbf{G}^\top \text{diag}(\text{vec}(\mathbf{W}))^2 \mathbf{G} + \rho \mathbf{I} \right) \text{vec}(\mathbf{Z}_m) \\ & = [\mathbf{I}_N \otimes \mathbf{R}]^\top \text{vec}(\mathbf{Y}_m) + \lambda_p \mathbf{G}^\top \text{diag}(\text{vec}(\mathbf{W}))^2 \mathbf{G} \text{vec}(\mathbf{Z}_h) \\ & + \rho \text{vec}(\mathbf{V}_m). \end{aligned} \quad (22)$$

Considering that the matrix on the left-hand side is symmetric positive-definite, the CG algorithm is used to solve this problem.

**Optimization w.r.t.  $\mathbf{V}_h$ :** This optimization problem can be written as:

$$\min_{\mathbf{V}_h} \frac{\rho}{2} \|\mathbf{V}_h - \mathbf{Z}_h\|_F^2 + \lambda_h \phi_h(\mathbf{V}_h). \quad (23)$$

As discussed above, designing accurate handcrafted regularizers for  $\phi_h(\mathbf{V}_h)$  may be complicated. To address this issue efficiently, we propose to use a strategy that leverages a CNN denoiser. Popular strategies are the Plug-and-Play (PnP) framework [46] and the Regularization by Denoising (RED) scheme [47]. In this work, we consider the RED strategy since it is associated with an explicit optimization objective and because it was experimentally shown in [47] to have more stable convergence and robustness in relation to the selection of hyperparameters when compared to PnP methods. Consider

denoising an HI  $\mathbf{V}$ , we define the CNN denoiser as  $\mathcal{D}(\mathbf{V})$ . RED framework defines  $\phi_h(\cdot)$  as the inner product between an image and its denoising residual:

$$\phi_h(\mathbf{V}) = \frac{1}{2} \langle \mathbf{V}, \mathbf{V} - \mathcal{D}(\mathbf{V}) \rangle, \quad (24)$$

where  $\langle \cdot, \cdot \rangle$  denotes the inner product. This can be interpreted as an image-adaptive Laplacian regularizer. Using (24), the optimization problem (23) becomes

$$\min_{\mathbf{V}_h} \frac{\rho}{2} \|\mathbf{V}_h - \mathbf{Z}_h\|_F^2 + \frac{\lambda_h}{2} \langle \mathbf{V}_h, \mathbf{V}_h - \mathcal{D}(\mathbf{V}_h) \rangle. \quad (25)$$

Taking the derivative of the cost function and setting it to zero, we obtain:

$$\rho(\mathbf{V}_h - \mathbf{Z}_h) + \lambda_h(\mathbf{V}_h - \mathcal{D}(\mathbf{V}_h)) = \mathbf{0}. \quad (26)$$

To solve this equation, a fixed-point iterative update is used, leading to the following recursive update equation:

$$\mathbf{V}_h^{(i+1)} = \frac{1}{\rho + \lambda_h} (\rho \mathbf{Z}_h + \lambda_h \mathcal{D}(\mathbf{V}_h^{(i)})). \quad (27)$$

where  $\mathbf{V}_h^{(i)}$  denotes the solution  $\mathbf{V}_h$  at the  $i$ -th iteration.

**Optimization w.r.t.  $\mathbf{V}_m$ :** Following the same strategy as above, we obtain:

$$\mathbf{V}_m^{(i+1)} = \frac{1}{\rho + \lambda_m} (\rho \mathbf{Z}_m + \lambda_m \mathcal{D}(\mathbf{V}_m^{(i)})). \quad (28)$$

Note that we only use a single step for the fixed point iteration in (27) and (28) for computational efficiency.

#### D. Learning deep priors via image-specific CNNs

Generally, function  $\mathcal{D}(\cdot)$  can be any off-the-shelf denoiser. This offers the opportunity of incorporating a fast CNN denoising engine with powerful prior learning ability into physical model-based iterative optimization procedure [27]. However, there are three main challenges in using CNN denoisers to learn priors for hyperspectral images in RED or PnP frameworks [28], [60]: First, there is a limited amount of data available for training; second, there is an even greater scarcity of labeled training data; third, the noise level of the HRI to be denoised in (27)-(28) changes over the BCD iterations as the method converges. To overcome each of these challenges, we propose a lightweight, unsupervised and image-specific CNN denoiser, which is detailed in the following.

**Lightweight network architecture:** To overcome the limited number of available data to train efficient CNN denoisers, a lightweight architecture with fewer parameters needs to be considered in the network design. In this work, two strategies have been considered to lighten network architecture, namely: 1) dimensionality reduction of the input image, which reduces the number of CNN filters, and 2) separable convolutions [61], which reduces the filter volume (i.e., the number of parameters of each filter).

We considered the DnCNN [62] as a backbone in network design. For color (i.e., RGB) images, each layer of DnCNN contains 64 filters. Directly using this network architecture to denoise an HI  $\mathbf{V}$  with  $L_h$  channels would approximately lead to the use of  $64 \times L_h/3$  filters in each layer, leading to a very

large number of parameters. This increase in the number of network parameters makes it hard to train since the amount of training data is usually very limited. Considering that the spectral channels of  $\mathbf{V}$  are highly correlated and contain highly redundant information, we can assume that there exists a subspace of dimension much lower than  $L_h$  which captures all the information of  $\mathbf{V}$ . This allows us to write  $\mathbf{V}$  using a low-rank representation as:

$$\mathbf{V} = \mathbf{Q}\mathbf{X}, \quad (29)$$

where  $\mathbf{Q} \in \mathbb{R}^{L_h \times l_h}$  ( $l_h \ll L_h$ ,  $\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}_{l_h}$ ) and  $\mathbf{X} \in \mathbb{R}^{l_h \times M}$  are the subspace matrix and the representation coefficients, respectively. Small values of  $l_h$  correspond to data description in a low-dimensional space. Employing such dimensionality reduction in the CNN denoising engine has a core benefit. It decreases the number of filters by a ratio of  $l_h/L_h$  in each layer by removing the burden of learning information that is redundant across spectral channels.

To reduce filter volume, we use separable convolutions to further lighten the backbone architecture as in [63]. In particular, the core idea of separable convolution is decomposing a convolution filter with  $3 \times 3 \times \text{Depth}$  parameters into a depth-wise filter with  $3 \times 3 \times 1$  parameters and a point-wise filter with  $1 \times 1 \times \text{Depth}$  parameters, where  $\text{Depth}$  is the input depth of this CNN layer. This reduces the number of parameters by a rate of  $1/\text{Depth} + 1/(3 \times 3)$ . Thus, the lightweight DnCNN contains three kinds of operators:  $3 \times 3$  separable convolution layers (S-Conv), rectified linear units (ReLU) and batch normalization (BN). ReLU is the activation function while BN is used to accelerate the training speed. In the network architecture, the first layer is ‘‘S-Conv + ReLU’’, the hidden layer is ‘‘S-Conv + BN + ReLU’’ and the last layer is ‘‘S-Conv’’. This network architecture is illustrated in the bottom panel of Figure 1. Furthermore, we adopt the residual learning strategy in [62] to predict the residual image before achieving the estimated clean image.

With these two strategies, the number of network parameters can be significantly reduced with a ratio of  $(l_h/L_h) \times (1/\text{Depth} + 1/(3 \times 3))$ , which is key to allowing the denoising engine to learn a powerful prior from a small training set.

**Zero-shot training strategy:** In many real-world scenarios, training data with paired noisy and clean images related to the scene of interest are not available. Moreover, using synthetic training data or images from different sites may lead to the so-called domain shift, where the model does not perform well due to differences between the statistical distribution of training and test data [28], [60]. Therefore, it is desirable to consider a training strategy that is *zero-shot*, that is, which is unsupervised and uses only the information of the observed noisy HI and MI pair itself for training.

Thus, we propose to leverage the information inside a single image to train the CNN denoiser. Natural images have significant information redundancy across different spatial positions and scales, which has been successfully exploited in single image restoration algorithms [64]. Consider the CNN-based denoiser  $\text{CNN}(\cdot; \Theta)$  with network parameters  $\Theta$ , and an observed noisy image  $\mathbf{X}$  generated following the degradation model  $\mathbf{X} = \mathbf{Z} + \mathbf{E}$ , where  $\mathbf{E}$  is i.i.d. Gaussian noise with a

---

**Algorithm 1** The Proposed CNN-based denoising engine.

---

**Input:** Noisy image  $\mathbf{V}$  and subspace dimension  $l_h$ .

**Output:** Denoised image  $\mathcal{D}(\mathbf{V})$ .

Find  $\mathbf{Q}$  and  $\mathbf{X}$  in (29) using the (truncated) SVD of  $\mathbf{V}$ .

Optimize  $\Theta$  by minimizing (30) with back-propagation.

Denoise  $\mathbf{X}$  with  $\Theta$  as  $\text{CNN}(\mathbf{X}; \Theta)$ .

Transform  $\text{CNN}(\mathbf{X}; \Theta)$  to  $\mathcal{D}(\mathbf{V}) = \mathbf{Q} \text{CNN}(\mathbf{X}; \Theta)$ .

---

standard deviation  $\sigma$ .  $\text{CNN}(\cdot; \Theta)$ . To learn the CNN denoiser  $\text{CNN}(\cdot; \Theta)$ , we make the important assumption that the set of parameters  $\Theta$  which allow it to recover  $\mathbf{Z}$  from  $\mathbf{X}$ , are the same as those which allow  $\text{CNN}(\cdot; \Theta)$  to recover  $\mathbf{X}$  from  $\mathbf{X} + \mathbf{E}$ . This assumption has been used to learn image-adapted CNNs for super-resolution in [48]. It allow us to train the denoising engine  $\text{CNN}(\cdot; \Theta)$  using the image pair  $(\mathbf{X} + \mathbf{E}, \mathbf{X})$  by minimizing the following  $\ell_1$ -norm loss function:

$$\ell(\Theta) = \|\text{CNN}(\mathbf{X} + \mathbf{E}; \Theta) - \mathbf{X}\|_1. \quad (30)$$

Note that the noisy-clean image pair  $(\mathbf{X} + \mathbf{E}, \mathbf{X})$  is generated by adding Gaussian noise with standard deviation  $\sigma$  to the observation  $\mathbf{X}$ . We adopted the method described in [65] to estimate  $\sigma$  in each channel of  $\mathbf{X}$ .

The procedure for learning the proposed CNN-based denoising engine is summarized in Algorithm 1. Note that the training procedure considers the entire image,  $\mathbf{X}$ . However, for large images, other learning objectives that decompose the image into different patches or across multiple scales can provide ways to parallelize the training procedure, which might reduce the execution times.

**Image-specific prior learning:** Since there exist some inter-image variations between  $\mathbf{Z}_h$  and  $\mathbf{Z}_m$ , we considered to train two independent denoising engines  $\text{CNN}(\cdot; \Theta_h)$  and  $\text{CNN}(\cdot; \Theta_m)$  to denoise  $\mathbf{V}_h$  and  $\mathbf{V}_m$ , respectively. This leads to different denoising engines, which can be expressed by substituting  $\mathcal{D}$  by  $\mathcal{D}_h$  in (27), and by  $\mathcal{D}_m$  in (28).

In general, the equivalent noise levels of  $\mathbf{V}_h$  and  $\mathbf{V}_m$  decrease over the BCD iterations since the reconstructed images get closer to the ground truth. Thus,  $\text{CNN}(\cdot; \Theta_h)$  and  $\text{CNN}(\cdot; \Theta_m)$  should have the ability to tackle multiple noise levels. To address this issue, we propose a strategy that adaptively updates network parameters  $\Theta_h$  and  $\Theta_m$  to learn an image-specific prior at each BCD iteration. This is performed by re-training  $\text{CNN}(\cdot; \Theta_h)$  and  $\text{CNN}(\cdot; \Theta_m)$  to denoise the estimates of the HRIs at the current BCD iteration. To make the algorithm faster, we consider training  $\text{CNN}(\cdot; \Theta_h)$  and  $\text{CNN}(\cdot; \Theta_m)$  in the first BCD iteration and then fine-tune them in all the remaining iterations.

Overall, after overcoming the discussed challenges with the above strategies, the denoising engine in Algorithm 1 is incorporated into the model-based optimization procedure described in Subsection III-C. The overall DIFIV strategy is described in Algorithm 2.

## IV. EXPERIMENTS

In this section, the effectiveness of the proposed DIFIV method is illustrated through numerical experiments consid-



**Algorithm 2** Deep Hyperspectral and Multispectral Image Fusion with Inter-image Variability (DIFIV).

**Input:**  $Y_h, Y_m, F, D, R$ , parameters  $p, \lambda_p, \lambda_h, \lambda_m, \rho$ .

**Output:** The estimated high-resolution images  $\hat{Z}_h, \hat{Z}_m$ .

Interpolate  $Y_h$  and  $Y_m$  as  $\tilde{Y}_h$  and  $\tilde{Y}_m$ , respectively.

Initialize  $Z_h = V_h = \tilde{Y}_h$  and  $Z_m = V_m = \tilde{Y}_m$ .

Initialize  $W$  using (12).

**while** stopping criteria are not met **do**

    Calculate  $Z_h$  by solving (20) via CG algorithm.

    Calculate  $Z_m$  by solving (22) via CG algorithm.

    Update  $W$  using (12).

    Learn deep priors via denoising  $V_h$  with Algorithm 1.

    Update  $V_h$  via (27).

    Learn deep priors via denoising  $V_m$  with Algorithm 1.

    Update  $V_m$  via (28).

**end while**

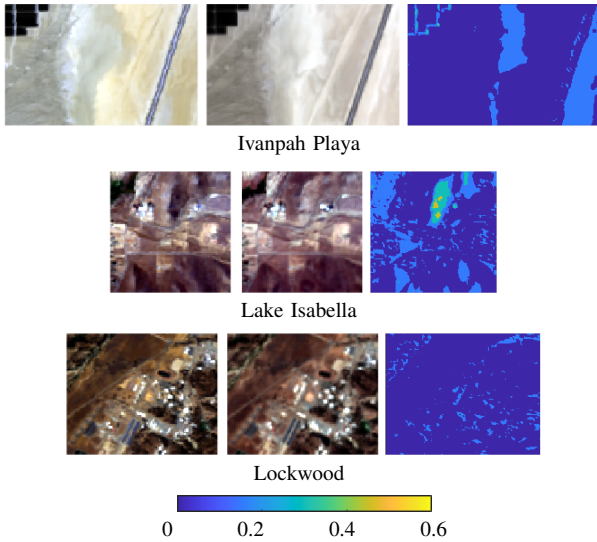


Figure 2. Visible representation of the hyperspectral (left panels) and multispectral images (middle panels) with moderate variability used in the experiments and their inter-image changes maps (right panels).

ering two categories of real data, i.e., observed images with moderate and significant inter-image variability. The results provided by the DIFIV are compared with other state-of-the-art hyperspectral and multispectral image fusion methods from both quantitative and qualitative perspectives. The code is made available at [https://github.com/xiuheng-wang/DIFIV\\_release](https://github.com/xiuheng-wang/DIFIV_release).

#### A. Experimental setup

We compared our method to nine other techniques, namely: the matrix factorization-based methods HySure [8] and CNMF [3], tensor-based image fusion methods STEREO [18] and SCOTT [20], the multiresolution analysis-based GLPHS algorithm [5], and the unsupervised deep learning based algorithm PAR [36]. We also considered approaches accounting for inter-image variability, including FuVar [37], GSFus [45] and CB-STAR [21]. In this study, three real data sets with moderate variability, namely, the Ivanpah Playa, the Lake

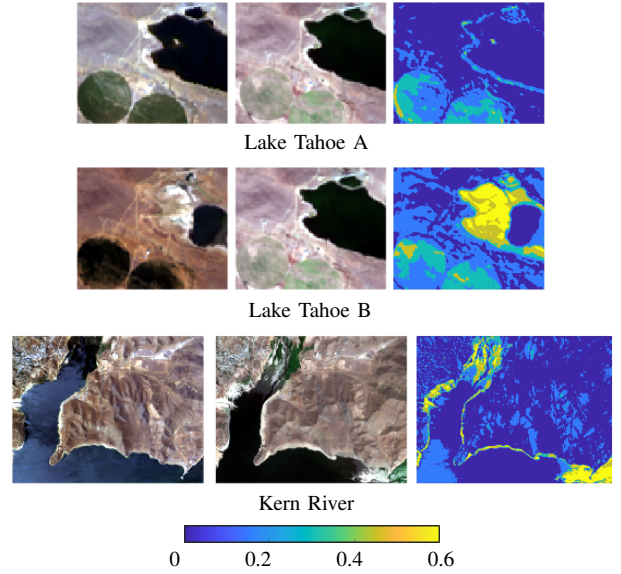


Figure 3. Visible representation of the hyperspectral (left panels) and multispectral images (middle panels) with significant variability used in the experiments and their inter-image changes maps (right panels).

Isabella and the Lookwood, and three real data sets with significant variability, namely, the Lake Tahoe A and B, and the Kern River, were used to evaluate the performance of each method. These data sets contained one reference HRI and an MI acquired by the AVIRIS and the Sentinel-2A instruments, respectively, with a pixel of 20m resolution [37]. The HI and MI contain  $L_h = 173$  and  $L_m = 10$  bands, respectively. To illustrate the existence of the inter-image variability in the considered datasets, we computed the average absolute difference images  $\frac{1}{L_m} \sum_{\ell=1}^{L_m} |Y_m(\ell, :) - R(\ell, :)Z_h|$  (where the modulus operation  $|\cdot|$  is applied elementwise), and displayed them in Figures 2 and 3.

For all acquired HRIs, which have the same spatial resolution as the MIs, a pre-processing procedure as described in [8] was performed. Specifically, spectral bands that were overly noisy or corresponded to water absorption spectral regions were removed manually, and then all bands of HRIs and MIs were normalized such that the 0.999 intensity quantile corresponded to the value of 1. Moreover, all HRIs were denoised using the approach described in [66] to obtain a noiseless reference image  $Z_h$ . The observed HRIs were generated according to (1), where  $F$  was an  $8 \times 8$  Gaussian blurring operator with standard deviation 4 and  $D$  a downsampling operator with the scaling factor 4. The SRF  $R$  was acquired from calibration measurements of the Sentinel-2A instrument and known a priori. For all experiments, Gaussian noise was added to both HRIs and MIs to obtain a signal-to-noise ratio (SNR) of 35 dB. To set up all baselines, we used the code provided by the authors and tuned all parameters to achieve the best fusion performance.

We implemented the proposed DIFIV method with the CNN-based denoising engine using the PyTorch framework. The dimension of subspace  $l_h$  was set to 5 and the number of network layers was set to 8, the first and hidden layers contained  $l_h \times 4$  S-Conv operators while the last layer was



composed by  $l_h$  S-Conv operators. The Adam optimizer [67] with an initial learning rate 0.0002 was used to minimize the loss function in (30). The number of iterations of DIFIV (Algorithm 2) was set to 20 which was sufficient to ensure convergence. The weights were initialized with the method in [68], trained for 10000 epochs in the first iteration, and fine-tuned for 2000 epochs in the remaining iterations. We set  $p = 1.5$ ,  $\lambda_p = 0.01$  and  $\lambda_m = \lambda_n = 0.1$  for the data with moderate variability. For the data with significant variability, we set  $p = 1.8$ ,  $\lambda_p = 0.002$  and  $\lambda_m = \lambda_n = 0.01$ . For the other parameters, we set  $\rho = 0.1$  and  $\epsilon = 10^{-6}$ . Note that in the following, the performance of the methods is compared via  $\mathbf{Z}_h$  only since the HRI corresponding to  $\mathbf{Y}_m$  was not available in the experiments.

### B. Quality measure and visual assessment

Four quality metrics were considered to evaluate the quality of the fusion result  $\hat{\mathbf{Z}}_h$  compared to the ground truth  $\mathbf{Z}_h$ . The first one is the peak signal to noise ratio (PSNR):

$$\text{PSNR} = \frac{1}{L_h} \sum_{\ell=1}^{L_h} 10 \log_{10} \left( \frac{M \max(\mathbf{Z}_h(\ell, :))^2}{\|\hat{\mathbf{Z}}_h(\ell, :) - \mathbf{Z}_h(\ell, :)\|^2} \right),$$

where  $\mathbf{Z}_h(\ell, :)$  and  $\hat{\mathbf{Z}}_h(\ell, :)$  represent the  $\ell$ -th channel of  $\mathbf{Z}_h$  and  $\hat{\mathbf{Z}}_h$ , respectively.

The second metric is the Spectral Angle Mapper (SAM):

$$\text{SAM} = \frac{1}{M} \sum_{m=1}^M \arccos \left( \frac{\hat{\mathbf{Z}}_h^\top(:, m) \mathbf{Z}_h(:, m)}{\|\hat{\mathbf{Z}}_h(:, m)\| \|\mathbf{Z}_h(:, m)\|} \right),$$

where  $\mathbf{Z}_h(:, m)$  and  $\hat{\mathbf{Z}}_h(:, m)$  denote the  $m$ -th pixel of  $\mathbf{Z}_h$  and  $\hat{\mathbf{Z}}_h$ , respectively.

The third metric is the ERGAS [69], which provides a global statistical measure of the fused image quality, defined as:

$$\text{ERGAS} = \frac{M}{N} \sqrt{\frac{10^4}{L_h} \sum_{\ell=1}^{L_h} \frac{\|\hat{\mathbf{Z}}_h(\ell, :) - \mathbf{Z}_h(\ell, :)\|^2}{\text{mean}(\hat{\mathbf{Z}}_h(\ell, :))^2}}.$$

This metric is the average of the UIQI [70] across bands. It evaluates image distortions including correlation loss and luminance and contrast distortions, and tends to 1 as  $\hat{\mathbf{Z}}_h$  tends to  $\mathbf{Z}_h$ .

For the visual assessment of the reconstructed images, we displayed color images at the visual spectrum (with band image at the wavelength 0.66, 0.56 and 0.45  $\mu m$  as red, green and blue channels) and false color images at the infrared spectrum (with band image at the wavelength 2.20, 1.50 and 0.80  $\mu m$  as red, green and blue channels). Due to space limitations, in the following, we only display the results of the five methods with the best quantitative performances, namely, CNMF, FuVar, GSFus, CB-STAR and DIFIV. Note that the last four algorithms account for inter-image variability.

### C. Category 1: Moderate variability

In this category, we evaluated the methods using HI and MI pairs with moderate variability, including Ivanpah Playa, Lake Isabella and Lockwood.

The first image pair considered in this category was acquired over the area surrounding Ivanpah Playa with a resolution of  $80 \times 128$  pixels. The second pair of images, with  $80 \times 80$  pixels, was captured over the Lake Isabella region, while the third pair of images containing  $80 \times 100$  pixels was acquired near Lockwood. The visualizations of these three image pairs and their inter-image variability are shown in Figure 2. In this category, the HI and MI look visually similar, which is typical when small differences between acquisition dates are considered (which is the case for the Lake Isabella and Lockwood images). Nevertheless, slight variations still exist, as can be seen in the overall color hue of the Ivanpah Playa and Lockwood images, and in the up part of the Lake Isabella image.

Table I  
RESULTS - IVANPAH PLAYA

Algorithm	SAM	ERGAS	PSNR	UIQI
HySure	2.262	2.639	21.923	0.511
CNMF	1.532	2.258	23.729	0.73
GLPHS	2.924	3.139	20.949	0.508
STEREO	29.173	1,643.756	17.744	0.49
SCOTT	41.025	618.314	9.388	0.307
PAR	3.506	2.26	24.011	0.752
FuVar	1.469	1.804	25.622	0.868
GSFus	1.72	1.497	27.264	0.874
CB-STAR	1.91	1.517	27.506	0.875
DIFIV	<b>1.358</b>	<b>1.335</b>	<b>28.283</b>	<b>0.903</b>

Table II  
RESULTS - LAKE ISABELLA

Algorithm	SAM	ERGAS	PSNR	UIQI
HySure	3.021	5.363	19.905	0.637
CNMF	2.206	3.414	25.611	0.792
GLPHS	2.755	3.572	25.207	0.793
STEREO	27.859	2,145.707	19.221	0.573
SCOTT	26.281	282.097	8.453	0.076
PAR	7.482	4.044	25.454	0.805
FuVar	2.487	3.234	27.213	0.899
GSFus	2.759	3.787	26.448	0.864
CB-STAR	3.263	3.406	26.556	0.864
DIFIV	<b>2.114</b>	<b>2.323</b>	<b>29.186</b>	<b>0.923</b>

Table III  
RESULTS - LOCKWOOD

Algorithm	SAM	ERGAS	PSNR	UIQI
HySure	3.384	4.384	22.678	0.881
CNMF	<b>3.243</b>	3.349	26.469	0.857
GLPHS	3.706	3.971	24.704	0.781
STEREO	28.185	883.508	21.079	0.639
SCOTT	20.109	204.538	9.273	0.094
PAR	6.61	4.433	23.634	0.754
FuVar	3.518	3.345	26.509	0.874
GSFus	3.331	3.332	26.329	0.87
CB-STAR	4.137	3.867	25.535	0.805
DIFIV	3.394	<b>2.934</b>	<b>27.307</b>	<b>0.885</b>

SAM, PSNR, ERGAS and UIQI metrics for all methods are reported in Table I to III. As shown in Table I and II, DIFIV outperforms all competing methods in all metrics for the Ivanpah Playa and Lake Isabella images. Moreover, it can

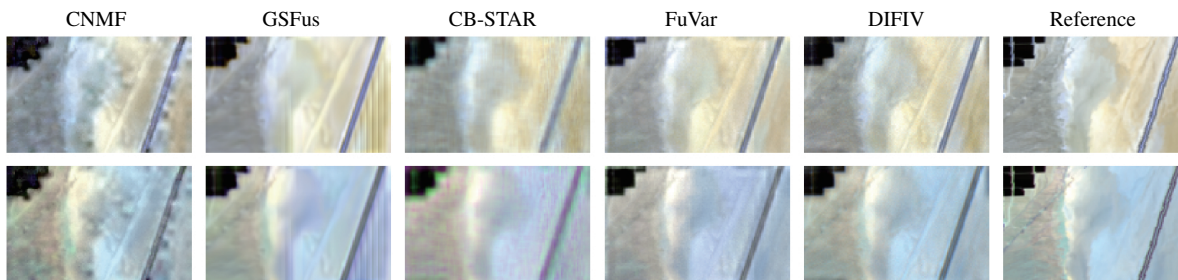


Figure 4. Visible (top) and infrared (bottom) representation for the estimated and true versions of the Ivanpah Playa HI.

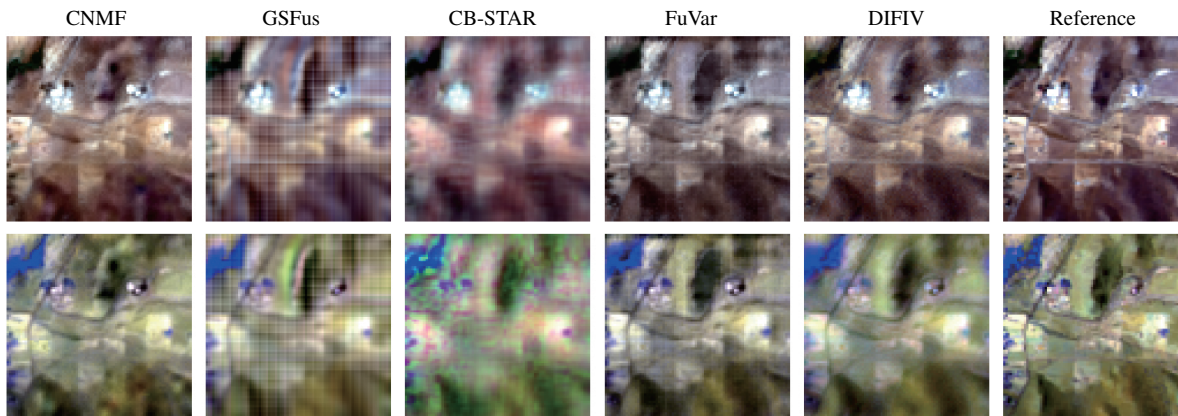


Figure 5. Visible (top) and infrared (bottom) representation for the estimated and true versions of the Lake Isabella HI.

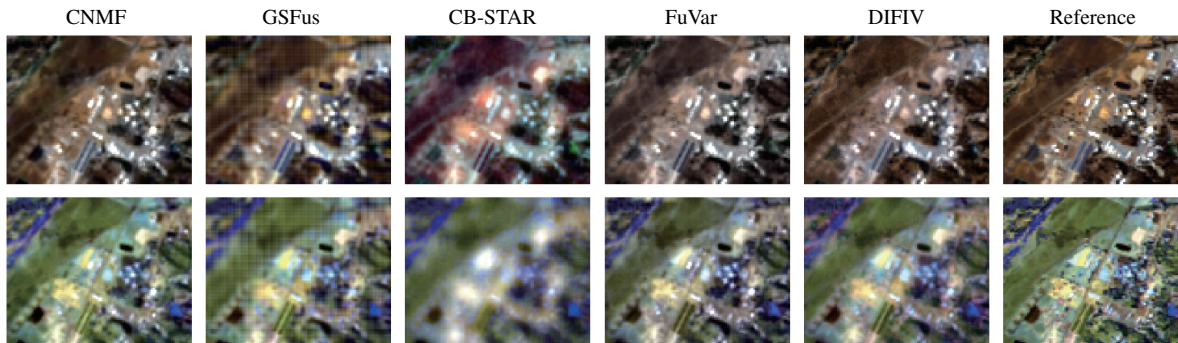


Figure 6. Visible (top) and infrared (bottom) representation for the estimated and true versions of the Lockwood HI.

be seen in Table III that DIFIV achieves overall best results for the Lookwood data, surpassing the other methods in all metrics except for SAM, where CNMF yields the best results for this metric. Figures 4 to 6 illustrate the color and false color visualization of the fusion results of several algorithms. Visually, DIFIV provides the best results in recovering details and spatial reconstructions closest to the ground truth at both the visible and infrared spectra. Specifically, CNMF and GSFus introduce artifacts and fail to recover many details while CB-STAR produces blurry effects and color aberrations. FuVar and DIFIV give similar visual effects but FuVar shows more details that do not match the reference image. This demonstrates the efficiency of DIFIV in recovering the spatial information of the latent HRIs in this category.

#### D. Category 2: Significant variability

This category evaluates the performance of the different methods when there is significant inter-image variability. We consider two image pairs acquired over the Lake Tahoe area at different time instant, namely, Lake Tahoe A and B. Besides, an image pair captured over the Kern River scene, which comprises a larger spatial area, was also considered.

The two Lake Tahoe image pairs contain  $100 \times 80$  pixels, while the Kern River image pair contains  $260 \times 340$  pixels. The visualization of these HIs and MIs and their corresponding inter-image changes maps can be seen in Figure 3. Significant variability between the HI and MI can be easily verified in these cases. For the two Lake Tahoe image pairs in this category, the color hue of the ground and the crop circles is quite different. Moreover, an island on the lake is not visible in the MI of Lake Tahoe A. For Lake Tahoe B, the lake in



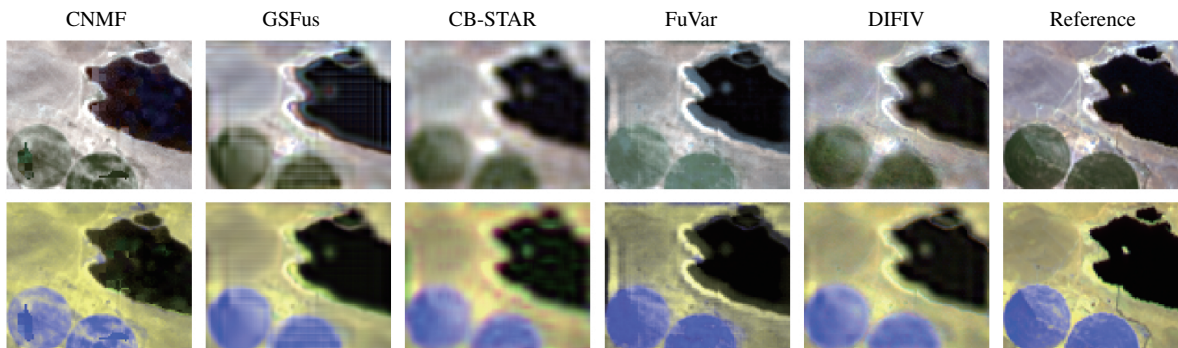


Figure 7. Visible (top) and infrared (bottom) representation for the estimated and true versions of the Lake Tahoe A HI.

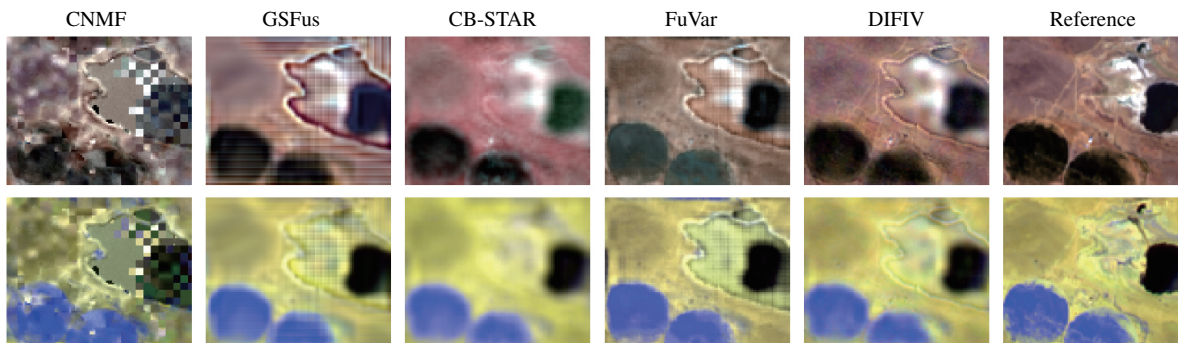


Figure 8. Visible (top) and infrared (bottom) representation for the estimated and true versions of the Lake Tahoe B HI.

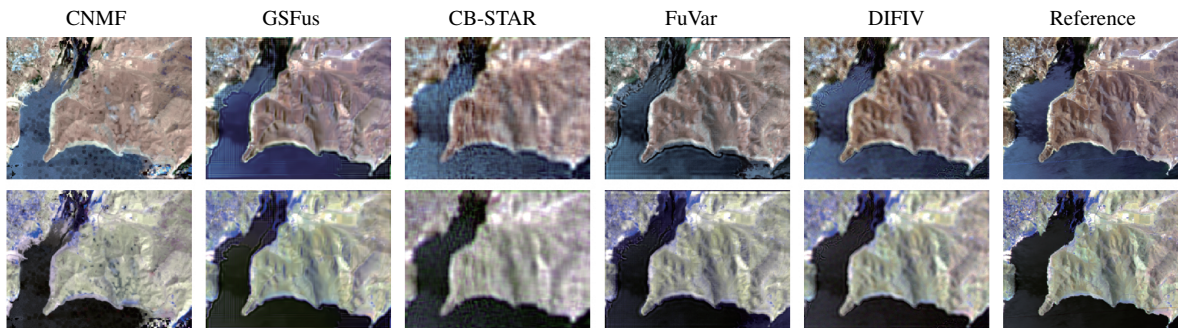


Figure 9. Visible (top) and infrared (bottom) representation for the estimated and true versions of the Kern River HI.

the MI is much larger than that in the HI. For the Kern River image pair, the river in the MI is narrower, has an upstream deposit, and shows a darker color in the water area.

The quantitative metrics are reported in Table IV, V and VI. As shown in Table IV, DIFIV obtains the best results for most metrics for Lake Tahoe A and only performs slightly worse in terms of SAM compared to GSFus. It can be observed in Table V and VI that the performance of DIFIV for Lake Tahoe B and Kern River exceeds those of the competing methods for all metrics. A visual illustration of the fusion results for Lake Tahoe A and B in color and false color is displayed in Figure 7 and Figure 8. Figure 9 shows the visualization of the fusion results for the Kern River dataset. It can be seen that DIFIV reconstructs more details and produces a color hue closer to the reference images at both visual and infrared spectral ranges. In particular, CNMF produced many artifacts and loses some details. GSFus and FuVar generate results with

blockiness and ghosting effects while the results of CB-STAR are blurry and have some color distortions. This demonstrates the superiority of DIFIV in recovering the latent HRIs when significant variability exists.

Table IV  
RESULTS - LAKE TAHOE A

Algorithm	SAM	ER GAS	PSNR	UIQI
HySure	10.643	7.775	16.531	0.655
CNMF	12.371	7.514	18.102	0.676
GLPHS	10.803	7.206	18.303	0.701
STEREO	30.605	2,541.149	15.991	0.575
SCOTT	42.839	457.101	9.243	0.215
PAR	15.886	6.065	20.579	0.811
FuVar	8.373	6.545	19.258	0.78
GSFus	<b>6.628</b>	4.376	22.537	0.883
CB-STAR	7.548	3.769	24.165	0.917
DIFIV	6.737	<b>3.706</b>	<b>24.174</b>	<b>0.922</b>

Table V  
RESULTS - LAKE TAHOE B

Algorithm	SAM	ERGAS	PSNR	UIQI
HySure	13.458	12.042	11.913	0.235
CNMF	7.954	7.289	16.387	0.428
GLPHS	6.662	4.786	19.824	0.665
STEREO	29.877	7,936.808	15.208	0.463
SCOTT	42.427	491.817	7.504	0.136
PAR	11.787	6.21	21.405	0.728
FuVar	4.688	3.729	21.86	0.79
GSFus	4.182	3.16	23.425	0.826
CB-STAR	3.95	2.597	25.221	0.881
DIFIV	<b>3.265</b>	<b>2.396</b>	<b>25.834</b>	<b>0.899</b>

Table VI  
RESULTS - KERN RIVER

Algorithm	SAM	ERGAS	PSNR	UIQI
HySure	9.094	8.933	21.717	0.442
CNMF	5.851	8.471	22.853	0.356
GLPHS	8.231	7.279	24.19	0.492
STEREO	30.337	636.136	22.568	0.45
SCOTT	27.652	220.14	13.239	0.045
PAR	11.695	6.742	28	0.739
FuVar	4.654	5.144	28.335	0.797
GSFus	5.037	4.243	29.404	0.785
CB-STAR	5.298	5.004	28.884	0.729
DIFIV	<b>3.412</b>	<b>3.734</b>	<b>31.506</b>	<b>0.852</b>

### E. Parameter Sensitivity

In this subsection, we study the sensitivity of DIFIV to the choice of values for regularization parameters  $\lambda_p$ ,  $\lambda_h$ ,  $\lambda_m$ . Considering the Ivanpah Playa scene as an example, we varied each parameter individually while keeping the remaining ones fixed at the values described in Subsection IV-A. The PSNR values of the fusion results as a function of the ratio  $\log_{10}(\lambda/\lambda_{opt})$  are shown in Figure 10, where  $\lambda_{opt}$  is the empirically selected value of the corresponding parameters. The PSNR values of two selected competing methods (CB-STAR and GSFus) are also shown for reference. It can be observed that varying parameters of DIFIV even by various orders of magnitude only leads to moderate variations of PSNR values, which are consistently higher than that of the competing methods. Moreover, the parameters of GSFus and CB-STAR were adjusted to provide the best performance in each example, and their performance would likewise degrade if their parameters move away from their optimal values, as discussed in the original works [45], [21]. This indicates the performance of DIFIV is not overly sensitive to the choice of regularization parameters.

### F. Computational cost

This experiment aims at comparing the computational cost of the algorithms accounting for inter-image variability. DIFIV was implemented using Python while the remaining methods were implemented using MATLAB. We conducted all the experiments on a computer with an Intel Core i7-10700 CPU, 32-GB random access memory and an NVIDIA Quadro P2200 GPU. The execution times of the algorithms for all the tested image pairs are shown in Table VII. It can be seen that the

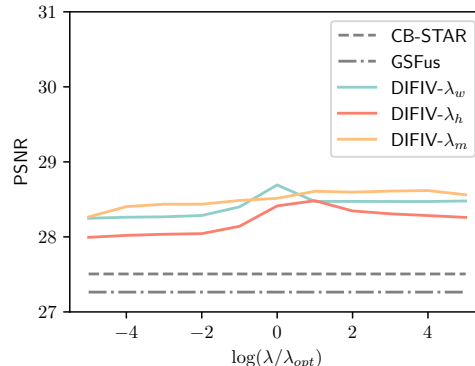


Figure 10. Sensitivity of the proposed DIFIV method with respect to regularization parameters  $\lambda_p$ ,  $\lambda_h$ ,  $\lambda_m$ .

Table VII  
EXECUTION TIMES OF THE ALGORITHMS THAT CONSIDER INTER-IMAGE VARIABILITY (IN SECONDS)

	FuVar	GSFus	CB-STAR	DIFIV
Ivanpah Playa	354.6	31.4	11.1	2963.4
Lake Isabella	199.3	17.7	8.8	2928.7
Lockwood	228.8	23.2	30.1	2954.0
Lake Tahoe A	679.5	23.1	7.8	2178.4
Lake Tahoe B	718.9	22.2	7.6	2143.5
Kern River	1762.0	307.3	96.0	5908.4

computation times of DIFIV are substantially higher than those of the competing methods, which comes as a compromise for its superior image fusion quality results. Nevertheless, the computation times of DIFIV scale reasonably with the image sizes; for instance, comparing the results for the Lake Isabella and Kern River images, we see that an increase of about ten times in the number of pixels in the image leads to an increase of about two times in the computation times. The development of computationally efficient extensions to the DIFIV method will be investigated in future work.

## V. CONCLUSIONS

This paper presented an unsupervised deep learning-based HMIF method accounting for inter-image variability. We first formulated a new imaging model considering both the joint as well as the image-specific priors related to the two latent HRIs. The inter-image variations were modeled using a hyper-Laplacian distribution, while the image-specific priors of the latent HRIs were defined implicitly by deep denoising engines. An iteratively reweighted scheme was then investigated to solve the non-convex cost function and tackle the joint image prior term. The optimization problem was solved using a variable splitting strategy, and the deep image priors were implemented by means of CNN-based denoising operations. A lightweight, image-specific CNN-based denoiser with a zero-shot training strategy was designed. The network parameters were iteratively updated during the optimization procedure in order to adapt to variations in the statistical properties of the estimated HRIs as the method converged. The proposed method achieved superior experimental performance in the presence of both moderate and significant inter-image variability when compared to state-of-the-art approaches.



## REFERENCES

- [1] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geoscience and Remote Sensing Magazine*, vol. 1, no. 2, pp. 6–36, 2013.
- [2] G. A. Shaw and H.-h. K. Burke, "Spectral imaging for remote sensing," *Lincoln Laboratory Journal*, vol. 14, no. 1, pp. 3–28, 2003.
- [3] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 2, pp. 528–537, 2012.
- [4] J. Liu, "Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details," *International Journal of Remote Sensing*, vol. 21, no. 18, pp. 3461–3472, 2000.
- [5] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "MTF-tailored multiscale fusion of high-resolution MS and pan imagery," *Photogrammetric Engineering & Remote Sensing*, vol. 72, no. 5, pp. 591–596, 2006.
- [6] G. Vivone, R. Restaino, and J. Chanussot, "Full scale regression-based injection coefficients for panchromatic sharpening," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3418–3431, 2018.
- [7] L. Loncan, L. B. de Almeida, J. M. Bioucas-Dias, X. Briottet, J. Chanussot, N. Dobigeon, S. Fabre, W. Liao, G. A. Licciardi, M. Simoes et al., "Hyperspectral pansharpening: A review," *IEEE Geoscience and remote sensing magazine*, vol. 3, no. 3, pp. 27–46, 2015.
- [8] M. Simões, J. Bioucas-Dias, L. B. Almeida, and J. Chanussot, "A convex formulation for hyperspectral image superresolution via subspace-based regularization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 6, pp. 3373–3388, 2015.
- [9] N. Keshava and J. F. Mustard, "Spectral unmixing," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 44–57, 2002.
- [10] N. Dobigeon, J.-Y. Tourneret, C. Richard, J. C. M. Bermudez, S. McLaughlin, and A. O. Hero, "Nonlinear unmixing of hyperspectral images: Models and algorithms," *IEEE Signal processing magazine*, vol. 31, no. 1, pp. 82–94, 2013.
- [11] Q. Wei, J. Bioucas-Dias, N. Dobigeon, and J.-Y. Tourneret, "Hyperspectral and multispectral image fusion based on a sparse representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 7, pp. 3658–3668, 2015.
- [12] N. Akhtar, F. Shafait, and A. Mian, "Bayesian sparse representation for hyperspectral image super resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3631–3640.
- [13] R. Kawakami, Y. Matsushita, J. Wright, M. Ben-Ezra, Y.-W. Tai, and K. Ikeuchi, "High-resolution hyperspectral imaging via matrix factorization," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Colorado Springs, CO, USA: IEEE, 2011, pp. 2329–2336.
- [14] C. Lanaras, E. Baltsavias, and K. Schindler, "Hyperspectral super-resolution by coupled spectral unmixing," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3586–3594.
- [15] M. A. Veganzones, M. Simoes, G. Licciardi, N. Yokoya, J. M. Bioucas-Dias, and J. Chanussot, "Hyperspectral super-resolution of locally low rank images from complementary multisource data," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 274–288, 2016.
- [16] Q. Wei, N. Dobigeon, and J.-Y. Tourneret, "Fast fusion of multi-band images based on solving a Sylvester equation," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4109–4121, 2015.
- [17] L. Zhang, W. Wei, C. Bai, Y. Gao, and Y. Zhang, "Exploiting clustering manifold structure for hyperspectral imagery super-resolution," *IEEE Transactions on Image Processing*, vol. 27, no. 12, pp. 5969–5982, 2018.
- [18] C. I. Kanatsoulis, X. Fu, N. D. Sidiropoulos, and W.-K. Ma, "Hyperspectral super-resolution: A coupled tensor factorization approach," *IEEE Transactions on Signal Processing*, vol. 66, no. 24, pp. 6503–6517, 2018.
- [19] S. Li, R. Dian, L. Fang, and J. M. Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4118–4130, 2018.
- [20] C. Prévost, K. Usevich, P. Comon, and D. Brie, "Hyperspectral super-resolution with coupled Tucker approximation: Recoverability and SVD-based algorithms," *IEEE Transactions on Signal Processing*, vol. 68, pp. 931–946, 2020.
- [21] R. A. Borsoi, C. Prévost, K. Usevich, D. Brie, J. C. M. Bermudez, and C. Richard, "Coupled tensor decomposition for hyperspectral and multispectral image fusion with inter-image variability," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 3, pp. 702–717, 2021.
- [22] M. Ding, X. Fu, T.-Z. Huang, J. Wang, and X.-L. Zhao, "Hyperspectral super-resolution via interpretable block-term tensor modeling," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 3, pp. 641–656, 2020.
- [23] J. Li, D. Hong, L. Gao, J. Yao, K. Zheng, B. Zhang, and J. Chanussot, "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *arXiv preprint arXiv:2205.01380*, 2022.
- [24] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, and Z. Xu, "Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution," in *European Conference on Computer Vision*. Springer, 2020, pp. 208–224.
- [25] L. Zhang, J. Nie, W. Wei, Y. Zhang, S. Liao, and L. Shao, "Unsupervised adaptation learning for hyperspectral imagery super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3073–3082.
- [26] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Multispectral and hyperspectral image fusion using a 3-D-convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 639–643, 2017.
- [27] J. Chen, M. Zhao, X. Wang, C. Richard, and S. Rahardja, "Integration of physics-based and data-driven models for hyperspectral image unmixing: A summary of current methods," *IEEE Signal Processing Magazine*, vol. 40, no. 2, pp. 61–74, 2023.
- [28] R. Dian, S. Li, and X. Kang, "Regularizing hyperspectral and multispectral image fusion by CNN denoiser," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 3, pp. 1124–1135, 2020.
- [29] X. Wang, J. Chen, Q. Wei, and C. Richard, "Hyperspectral image super-resolution via deep prior regularization with parameter estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 1708–1723, 2021.
- [30] Q. Xie, M. Zhou, Q. Zhao, Z. Xu, and D. Meng, "MHF-net: An interpretable deep network for multispectral and hyperspectral image fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [31] Y. Qu, H. Qi, and C. Kwan, "Unsupervised sparse Dirichlet-net for hyperspectral image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2511–2520.
- [32] J. Liu, Z. Wu, L. Xiao, and X.-J. Wu, "Model inspired autoencoder for unsupervised hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, 2022.
- [33] Z. Wang, B. Chen, R. Lu, H. Zhang, H. Liu, and P. K. Varshney, "FusionNet: An unsupervised convolutional variational network for hyperspectral and multispectral image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 7565–7577, 2020.
- [34] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9446–9454.
- [35] L. Zhang, J. Nie, W. Wei, Y. Li, and Y. Zhang, "Deep blind hyperspectral image super-resolution," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 6, pp. 2388–2400, 2020.
- [36] W. Wei, J. Nie, L. Zhang, and Y. Zhang, "Unsupervised recurrent hyperspectral imagery super-resolution using pixel-aware refinement," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2020.
- [37] R. A. Borsoi, T. Imbiriba, and J. C. M. Bermudez, "Super-resolution for hyperspectral and multispectral image fusion accounting for seasonal spectral variability," *IEEE Transactions on Image Processing*, vol. 29, no. 1, pp. 116–127, 2020.
- [38] N. Yokoya, C. Grohnfeldt, and J. Chanussot, "Hyperspectral and multispectral data fusion: A comparative review of the recent literature," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 2, pp. 29–56, 2017.
- [39] R. A. Borsoi, T. Imbiriba, J. C. M. Bermudez, C. Richard, J. Chanussot, L. Drumetz, J.-Y. Tourneret, A. Zare, and C. Jutten, "Spectral variability in hyperspectral data unmixing: A comprehensive review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 4, pp. 223–270, 2021.
- [40] S. Liu, D. Marinelli, L. Bruzzone, and F. Bovolo, "A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 2, pp. 140–158, 2019.
- [41] C. Prévost, R. A. Borsoi, K. Usevich, D. Brie, J. C. M. Bermudez, and C. Richard, "Hyperspectral super-resolution accounting for spectral variability: Coupled tensor LL1-based recovery and blind unmixing of the unknown super-resolution image," *SIAM Journal on Imaging Sciences*, vol. 15, no. 1, pp. 110–138, 2022.
- [42] R. A. Borsoi, C. Prévost, K. Usevich, D. Brie, J. C. M. Bermudez, and C. Richard, "Coupled tensor models accounting for inter-image

- variability,” in *2021 55th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2021, pp. 1586–1590.
- [43] S. E. Brezini, M. S. Karoui, F. Z. Benhalouche, Y. Deville, and A. Ouamri, “Hypersharpener by an NMF-unmixing-based method addressing spectral variability,” *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.
- [44] A. Camacho, E. Vargas, and H. Arguello, “Hyperspectral and multispectral image fusion addressing spectral variability by an augmented linear mixing model,” *International Journal of Remote Sensing*, vol. 43, no. 5, pp. 1577–1608, 2022.
- [45] X. Fu, S. Jia, M. Xu, J. Zhou, and Q. Li, “Fusion of hyperspectral and multispectral images accounting for localized inter-image changes,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2021.
- [46] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, “Plug-and-play priors for model based reconstruction,” in *2013 IEEE Global Conference on Signal and Information Processing*. IEEE, 2013, pp. 945–948.
- [47] Y. Romano, M. Elad, and P. Milanfar, “The little engine that could: Regularization by denoising (RED),” *SIAM Journal on Imaging Sciences*, vol. 10, no. 4, pp. 1804–1844, 2017.
- [48] A. Shocher, N. Cohen, and M. Irani, ““zero-shot” super-resolution using deep internal learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3118–3126.
- [49] T. Imbiriba, R. A. Borsoi, and J. C. M. Bermudez, “Generalized linear mixing model accounting for endmember variability,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 1862–1866.
- [50] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, “Tensor decomposition for signal processing and machine learning,” *IEEE Transactions on Signal Processing*, vol. 65, no. 13, pp. 3551–3582, 2017.
- [51] C. Prévost, K. Usevich, P. Comon, and D. Brie, “Coupled tensor low-rank multilinear approximation for hyperspectral super-resolution,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, U.K.: IEEE, 2019, pp. 5536–5540.
- [52] X. Wang, J. Chen, and C. Richard, “Hyperspectral image super-resolution with deep priors and degradation model inversion,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 2814–2818.
- [53] D. Krishnan and R. Fergus, “Fast image deconvolution using hyper-laplacian priors,” *Advances in Neural Information Processing Systems*, vol. 22, pp. 1033–1041, 2009.
- [54] Y. Peng, W. Li, X. Luo, and J. Du, “Hyperspectral image superresolution using global gradient sparse and nonlocal low-rank tensor decomposition with hyper-laplacian prior,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 5453–5469, 2021.
- [55] R. A. Borsoi, G. H. Costa, and J. C. M. Bermudez, “A new adaptive video super-resolution algorithm with improved robustness to innovations,” *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 673–686, 2018.
- [56] Z. Lu, “Iterative reweighted minimization methods for  $\ell_p$  regularized unconstrained nonlinear programming,” *Mathematical Programming*, vol. 147, no. 1, pp. 277–307, 2014.
- [57] R. Ammanouil, A. Ferrari, C. Richard, and D. Mary, “Blind and fully constrained unmixing of hyperspectral images,” *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5510–5518, 2014.
- [58] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, “Iteratively reweighted least squares minimization for sparse recovery,” *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 63, no. 1, pp. 1–38, 2010.
- [59] D. Geman and C. Yang, “Nonlinear image recovery with half-quadratic regularization,” *IEEE Transactions on Image Processing*, vol. 4, no. 7, pp. 932–946, 1995.
- [60] X. Wang, J. Chen, and C. Richard, “Tuning-free plug-and-play hyperspectral image deconvolution with deep priors,” *IEEE Transactions on Geoscience and Remote Sensing*, 2023.
- [61] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861*, 2017.
- [62] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, “Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising,” *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.
- [63] R. Imamura, T. Itasaka, and M. Okuda, “Zero-shot hyperspectral image denoising with separable image prior,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.
- [64] D. Glasner, S. Bagon, and M. Irani, “Super-resolution from a single image,” in *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 349–356.
- [65] D. L. Donoho and J. M. Johnstone, “Ideal spatial adaptation by wavelet shrinkage,” *biometrika*, vol. 81, no. 3, pp. 425–455, 1994.
- [66] R. E. Roger and J. F. Arnold, “Reliably estimating the noise in AVIRIS hyperspectral images,” *International Journal of Remote Sensing*, vol. 17, no. 10, pp. 1951–1962, 1996.
- [67] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [68] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.
- [69] L. Wald, “Quality of high resolution synthesised images: Is there a simple criterion?” in *Third conference Fusion of Earth data: merging point measurements, raster maps and remotely sensed images*. SEE/URISCA, 2000, pp. 99–103.
- [70] Z. Wang and A. C. Bovik, “A universal image quality index,” *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002.