Deep Image Fusion Accounting for Inter-Image Variability

Xiuheng Wang, Ricardo Augusto Borsoi, Cédric Richard, Jie Chen

DEEP IMAGE FUSION ACCOUNTING FOR INTER-IMAGE VARIABILITY

*Xiuheng Wang *, Ricardo Augusto Borsoi †, Cédric Richard *, Jie Chen ‡*

* Université Côte d'Azur, CNRS, OCA, France

† Université de Lorraine, CNRS, CRAN, France

‡ Research and Development Institute of Northwestern Polytechnical University in Shenzhen, China

xiuheng.wang@oca.eu, raborsoi@gmail.com, cedric.richard@unice.fr, dr.jie.chen@ieee.org

ABSTRACT

Hyperspectral and multispectral image fusion (HMIF) allows us to overcome inherent hardware limitations of hyperspectral imaging systems with respect to their lower spatial resolution. However, existing algorithms fail to consider realistic image acquisition conditions, or to leverage the powerful representation capacity of deep neural networks. This paper introduces a general imaging model which considers inter-image variability of data from heterogeneous sources, and formulates the optimization problem. Then it presents a new 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, leverages unsupervised light-weight CNN-based denoisers to learn realistic image priors from data. Its performance is illustrated with real data that suffer from inter-image variability.

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

1. INTRODUCTION

Hyperspectral imaging systems are able to record the reflectance of a scene in hundreds of narrow, contiguous spectral bands. Their rich spectral information has attracted high interest in the remote sensing literature, with successful applications in mineral exploration, vegetation monitoring and land cover analysis [1]. Nevertheless, the high spectral resolution of hyperspectral images (HI) restricts their spatial resolution due to hardware limitations. In contrast, multispectral cameras can achieve a much higher spatial resolution by only acquiring a small number of spectral bands. Consequently, a strategy of improving the spatial resolution of HIs is to fuse them with multispectral images (MI) of the same scene, resulting in the so-called HMIF problem.

This work has been was supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002. The work of J. Chen was supported in part by Guangdong International Cooperation Project 2022A0505050020, Shenzhen Research Grant JCYJ20220530161606014, Shaanxi Key Industrial Innovation Chain Project 2022ZDLGY01-02, and Xi'an Technology Industrialization Plan XA2020-RGZNTJ-0076.

Fusing HIs and MIs acquired by different instruments at different time instants remains a challenging problem [2]. Most methods ignore that there may have differences between the acquisition conditions of the HI and MI, which can compromise their performance. To tackle this issue, HMIF frameworks addressing inter-image variability have been proposed [3–5]. These methods formulate and solve optimization problems containing different high-resolution image (HRI) priors. Recently, learning priors from data with the Plug-and-Play (PnP) and the Regularization by Denoising (RED) [6] frameworks has become widely used in image processing because of their tractability and superior performance. In particular, convolutional neural networks (CNN) with deep architectures provide very efficient image priors [7]. However, existing HMIF algorithms either relied on handcrafted priors for the HRIs, or neglected to account for the joint and for the image-specific information when variability is present.

In this paper, we propose a new image fusion method accounting for inter-image variability between HIs and MIs. A general imaging model is formulated, where the joint prior of two HRIs is investigated. Specifically, the smoothness of inter-image variations is represented with an hyper-Laplacian distribution while the characteristics of each HRI are learned by CNNs. The HMIF optimization problem is solved by combining variable splitting with an iteratively reweighted scheme to deal with non-convex image priors. Considering the RED paradigm and its bottlenecks, an unsupervised light-weight CNN is specifically designed and then incorporated into the iterative optimization algorithm as a denoiser to learn priors of the two HRIs. 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.

2. GENERAL IMAGING MODEL

Let us denote by $\boldsymbol{Y}_h \in \mathbb{R}^{L_h \times N}$ and by $\boldsymbol{Y}_m \in \mathbb{R}^{L_m \times M}$ the observed HI and MI, respectively. These images are assumed to be degraded versions of a pair of HRIs $\boldsymbol{Z}_h \in \mathbb{R}^{L_h \times M}$ and

$\boldsymbol{Z}_m \in \mathbb{R}^{L_h \times M}$ as follows:

$$\boldsymbol{Y}_h = \boldsymbol{Z}_h \boldsymbol{F} \boldsymbol{D} + \boldsymbol{E}_h\,, \quad \boldsymbol{Y}_m = \boldsymbol{R}\boldsymbol{Z}_m + \boldsymbol{E}_m\,, \tag{1}$$

where matrices $\boldsymbol{F} \in \mathbb{R}^{M\times M}$, $\boldsymbol{D} \in \mathbb{R}^{M\times N}$ and $\boldsymbol{R} \in \mathbb{R}^{L_m \times L_h}$ represent optical blurring, spatial down-sampling, and the spectral response function (SRF) of the MI, respectively. $\boldsymbol{E}_h \in \mathbb{R}^{L_h \times N}$ and $\boldsymbol{E}_m \in \mathbb{R}^{L_m \times M}$ denote additive noises.

In this setting, the image fusion problem consists of recovering the HRIs $\boldsymbol{Z}_h$ and $\boldsymbol{Z}_m$ given the observations $\boldsymbol{Y}_h$ and $\boldsymbol{Y}_m$. Most of the previous methods consider that $\boldsymbol{Y}_h$ and $\boldsymbol{Y}_m$ are degraded from the same source, i.e., $\boldsymbol{Z}_h = \boldsymbol{Z}_m$, 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 [2]. In that case, by assuming that $\boldsymbol{Z}_h = \boldsymbol{Z}_m$, 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 [8], or abrupt changes [9].

3. THE PROPOSED METHOD

In a probabilistic framework, HMIF can be performed by maximizing the posterior probability distribution function (PDF) of the HRIs given the HI and MI:

$$p(\boldsymbol{Z}_h, \boldsymbol{Z}_m | \boldsymbol{Y}_h, \boldsymbol{Y}_m) \propto p(\boldsymbol{Y}_m, \boldsymbol{Y}_h | \boldsymbol{Z}_h, \boldsymbol{Z}_m) p(\boldsymbol{Z}_m, \boldsymbol{Z}_h)\,.$$

The main challenge is defining the prior $p(\boldsymbol{Z}_m, \boldsymbol{Z}_h)$, which should: 1) favor images $\boldsymbol{Z}_m$ and $\boldsymbol{Z}_h$ that are statistically similar to real hyperspectral images, and 2) introduce changes between $\boldsymbol{Z}_m$ and $\boldsymbol{Z}_h$ which, apart from possible smooth inter-image variations, are sparse. To achieve this desiderata, we consider the following prior:

$$\begin{aligned}\log p(\boldsymbol{Z}_m, \boldsymbol{Z}_h) \propto\; & -\frac{\lambda}{2}\sum_{\ell,n} \left|\delta_h^{(\ell,n)} - \delta_m^{(\ell,n)}\right|^p \\ & - \lambda_m \phi_m(\boldsymbol{Z}_m) - \lambda_h \phi_h(\boldsymbol{Z}_h)\,,\end{aligned} \tag{2}$$

where $\delta_h^{(\ell,n)}$ and $\delta_m^{(\ell,n)}$ denote the (ℓ, n)-th locations of a high-pass spatial-spectral filtered version of $\boldsymbol{Z}_h$ and $\boldsymbol{Z}_m$, denoted by $\boldsymbol{\Delta}_h = \mathcal{G}(\boldsymbol{Z}_h)$ and $\boldsymbol{\Delta}_m = \mathcal{G}(\boldsymbol{Z}_m)$, where $\mathcal{G}$ is the Laplacian operator. The spatial and spectral priors on $\boldsymbol{Z}_m$ and $\boldsymbol{Z}_h$ are encoded in $\phi(\boldsymbol{Z}_h)$ and $\phi(\boldsymbol{Z}_m)$, respectively. Parameter p is an exponent to be set, and λ_h, λ_m and λ are regularization parameters.

Assuming $\boldsymbol{E}_h$ and $\boldsymbol{E}_m$ to be jointly i.i.d. Gaussian, maximizing $p(\boldsymbol{Z}_h, \boldsymbol{Z}_m | \boldsymbol{Y}_h, \boldsymbol{Y}_m)$ in model (1) is equivalent to:

$$\begin{aligned}\min_{\boldsymbol{Z}_h, \boldsymbol{Z}_m}\; & \frac{1}{2}\|\boldsymbol{Y}_h - \boldsymbol{Z}_h\boldsymbol{F}\boldsymbol{D}\|_F^2 + \frac{1}{2}\|\boldsymbol{Y}_m - \boldsymbol{R}\boldsymbol{Z}_m\|_F^2 + \lambda_h\phi(\boldsymbol{Z}_h) \\ & + \lambda_m\phi(\boldsymbol{Z}_m) + \frac{\lambda}{2}\sum_{\ell,n}\left|\delta_h^{(\ell,n)} - \delta_m^{(\ell,n)}\right|^p\end{aligned} \tag{3}$$

3.1. An iteratively reweighted update scheme

To optimize inter-image prior term $\sum_{\ell,n}\left|\delta_h^{(\ell,n)} - \delta_m^{(\ell,n)}\right|^p$ which is non-convex and non-smooth, we consider an iteratively reweighted optimization strategy [10]. We propose to solve (3) by repeating the following steps until convergence:

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

$$\begin{aligned}\min_{\boldsymbol{Z}_h, \boldsymbol{Z}_m}\; & \frac{1}{2}\|\boldsymbol{Y}_h - \boldsymbol{Z}_h\boldsymbol{F}\boldsymbol{D}\|_F^2 + \frac{1}{2}\|\boldsymbol{Y}_m - \boldsymbol{R}\boldsymbol{Z}_m\|_F^2 + \lambda_h\phi(\boldsymbol{Z}_h) \\ & + \lambda_m\phi(\boldsymbol{Z}_m) + \frac{\lambda}{2}\|\boldsymbol{W}\odot(\boldsymbol{\Delta}_h - \boldsymbol{\Delta}_m)\|_F^2\,,\end{aligned} \tag{4}$$

where $\boldsymbol{W}$, $\boldsymbol{\Delta}_h$ and $\boldsymbol{\Delta}_m$ are matrices whose (ℓ, n)-th entries are given by $\sqrt{w_{\ell,n}}$, $\delta_h^{(\ell,n)}$ and $\delta_m^{(\ell,n)}$, respectively. Operator $\odot$ denotes the Hadamard product.

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

$$w_{\ell,n} = \left(|\delta_h^{(\ell,n)} - \delta_m^{(\ell,n)}| + \epsilon\right)^{p-2}\,, \tag{5}$$

where $\epsilon > 0$ is a small constant included to ensure the numerical stability of the algorithm. In the following, we focus on the resolution of the optimization problem (4).

3.2. The optimization problem

Introducing two auxiliary variables $\boldsymbol{V}_h$ and $\boldsymbol{V}_m$, the data fidelity and regularization terms can be decoupled by writing the augmented Lagrangian of this cost function (4) as:

$$\begin{aligned}\mathcal{L}_\rho(\boldsymbol{Z}_h, \boldsymbol{Z}_m, \boldsymbol{V}_h, \boldsymbol{V}_m) = & \frac{1}{2}\|\boldsymbol{Y}_h - \boldsymbol{Z}_h\boldsymbol{F}\boldsymbol{D}\|_F^2 \\ & + \frac{1}{2}\|\boldsymbol{Y}_m - \boldsymbol{R}\boldsymbol{Z}_m\|_F^2 + \frac{\lambda}{2}\|\boldsymbol{W}\odot(\boldsymbol{\Delta}_h - \boldsymbol{\Delta}_m)\|_F^2 \\ & + \frac{\rho}{2}\|\boldsymbol{Z}_m - \boldsymbol{V}_m\|_F^2 + \frac{\rho}{2}\|\boldsymbol{Z}_h - \boldsymbol{V}_h\|_F^2 \\ & + \lambda_m\phi(\boldsymbol{V}_m) + \lambda_h\phi(\boldsymbol{V}_h)\end{aligned} \tag{6}$$

where ρ is the penalty parameter. We minimize the cost function $\mathcal{L}_\rho$ with respect to each of its variables:

$$\begin{aligned}\min_{\boldsymbol{Z}_h}\; & \frac{1}{2}\|\boldsymbol{Y}_h - \boldsymbol{Z}_h\boldsymbol{F}\boldsymbol{D}\|_F^2 + \frac{\lambda}{2}\|\boldsymbol{W}\odot(\boldsymbol{\Delta}_h - \boldsymbol{\Delta}_m)\|_F^2 \\ & + \frac{\rho}{2}\|\boldsymbol{Z}_h - \boldsymbol{V}_h\|^2\,,\end{aligned} \tag{7}$$

$$\begin{aligned}\min_{\boldsymbol{Z}_m}\; & \frac{1}{2}\|\boldsymbol{Y}_m - \boldsymbol{R}\boldsymbol{Z}_m\|_F^2 + \frac{\lambda}{2}\|\boldsymbol{W}\odot(\boldsymbol{\Delta}_h - \boldsymbol{\Delta}_m)\|_F^2 \\ & + \frac{\rho}{2}\|\boldsymbol{V}_m - \boldsymbol{Z}_m\|^2\,,\end{aligned} \tag{8}$$

$$\min_{\boldsymbol{V}_h}\; \frac{\rho}{2}\|\boldsymbol{V}_h - \boldsymbol{Z}_h\|^2 + \lambda_h\phi(\boldsymbol{V}_h)\,, \tag{9}$$

$$\min_{\boldsymbol{V}_m}\; \frac{\rho}{2}\|\boldsymbol{V}_m - \boldsymbol{Z}_h\|^2 + \lambda_m\phi(\boldsymbol{V}_m)\,. \tag{10}$$

We propose to solve sub-problems (7) and (8) with the conjugate gradient (CG) algorithm. Because designing efficient regularizers $\phi(\boldsymbol{V}_h)$ and $\phi(\boldsymbol{V}_m)$ may be difficult, we propose

to use the RED [6] strategy, which leverages a powerful CNN-based denoiser $\mathcal{D}$, to solve sub-problems (9) and (10) as:

$$\boldsymbol{V}_h^{(i+1)} = \frac{1}{\rho + \lambda_h} \left(\rho \boldsymbol{Z}_h + \lambda_h \mathcal{D}(\boldsymbol{V}_h^{(i)}) \right), \tag{11}$$

$$\boldsymbol{V}_m^{(i+1)} = \frac{1}{\rho + \lambda_m} \left(\rho \boldsymbol{Z}_m + \lambda_m \mathcal{D}(\boldsymbol{V}_m^{(i)}) \right). \tag{12}$$

where $\boldsymbol{V}_h^{(i)}$ and $\boldsymbol{V}_m^{(i)}$ denote the solution $\boldsymbol{V}_h$ and $\boldsymbol{V}_m$ at the i-th iteration, respectively.

3.3. Learning deep prior via CNN

In our RED-based framework, three bottlenecks restrict the use of CNNs as efficient denoising engines for HIs: *limited amounts of training data*, *lack of labels*, and *multiple noise levels*. We propose to overcome these bottlenecks point by point with the following strategies.

Light-weight network architecture: To overcome the limited amount of data available to train CNN denoisers, two strategies were considered to design an architecture with few parameters, namely: 1) dimensionality reduction, and 2) separable convolutions [11]. We considered the DnCNN [12] as a backbone in network design. Considering that the spectral channels of an HI $\boldsymbol{V}$ contain highly redundant information, we assume that there exists a subspace of dimension l_h (much lower than L_h) which captures all the information of $\boldsymbol{V}$. This allows us to write $\boldsymbol{V}$ using a low-rank representation as:

$$\boldsymbol{V} = \boldsymbol{Q}\boldsymbol{X}, \tag{13}$$

where $\boldsymbol{Q} \in \mathbb{R}^{L_h \times l_h}$ (satisfying $\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}_{l_h}$) and $\boldsymbol{X} \in \mathbb{R}^{l_h \times M}$ are the subspace matrix and the representation coefficients, respectively. This decreases the number of filters by a ratio of l_h/L_h in each layer.

To reduce filter volume and further lighten the backbone architecture, we use separable convolutions as in [13]. The core idea is to decompose a convolution filter with $3 \times 3 \times \texttt{Depth}$ parameters into a depth-wise filter with $3 \times 3 \times 1$ parameters and a point-wise filter with $1 \times 1 \times \texttt{Depth}$ parameters, where $\texttt{Depth}$ is the input depth of this CNN layer. This reduces the number of parameters by a rate of $1/\texttt{Depth} + 1/(3 \times 3)$. Thus, the light-weight DnCNN contains three kinds of operators: 3×3 separable convolution layers (S-Conv), rectified linear units (ReLU) and batch normalization (BN). 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". With these two strategies, the number of network parameters can be reduced by a ratio of $(l_h/L_h) \times (1/\texttt{Depth} + 1/(3 \times 3))$.

Zero-shot training strategy: In many real-world scenarios, training data with paired noisy and clean images are not available. Moreover, using synthetic training data may lead to the domain shift [14,15]. Therefore, it is desirable to consider a training strategy that is both *unsupervised* and *zero-shot*, requiring only the observed HI and MI.

Algorithm 1 The Proposed CNN-based denoising engine.

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

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

Find $\boldsymbol{Q}$ and $\boldsymbol{X}$ in (13) using the (truncated) SVD of $\boldsymbol{V}$.
Optimize Θ by minimizing (14) with back-propagation.
Denoise $\boldsymbol{X}$ with Θ as $\texttt{CNN}(\boldsymbol{X}; \Theta)$.
Transform $\texttt{CNN}(\boldsymbol{X}; \Theta)$ to $\mathcal{D}(\boldsymbol{V}) = \boldsymbol{Q}\,\texttt{CNN}(\boldsymbol{X}; \Theta)$.

Algorithm 2 DIFIV.

Input: $\boldsymbol{Y}_h, \boldsymbol{Y}_m, \boldsymbol{F}, \boldsymbol{D}, \boldsymbol{R}$ paramters $p, \lambda, \lambda_h, \lambda_m, \rho$.

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

Interpolate $\boldsymbol{Y}_h$ and $\boldsymbol{Y}_m$ as $\widetilde{\boldsymbol{Y}}_h$ and $\widetilde{\boldsymbol{Y}}_m$, respectively.
Initialize $\boldsymbol{Z}_h = \boldsymbol{V}_h = \widetilde{\boldsymbol{Y}}_h$ and $\boldsymbol{Z}_m = \boldsymbol{V}_m = \widetilde{\boldsymbol{Y}}_m$.
Initialize $\boldsymbol{W}$ using (5).
while stopping criteria are not met **do**
  Calculate $\boldsymbol{Z}_h$ by solving (7) via CG algorithm.
  Calculate $\boldsymbol{Z}_m$ by solving (8) via CG algorithm.
  Update $\boldsymbol{W}$ using (5).
  Learn deep priors via denoising $\boldsymbol{V}_h$ with Algorithm 1.
  Update $\boldsymbol{V}_h$ via (11).
  Learn deep priors via denoising $\boldsymbol{V}_m$ with Algorithm 1.
  Update $\boldsymbol{V}_m$ via (12).
end while

We propose to leverage the information inside a single image to train the CNN denoiser. Consider the CNN-based denoiser $\texttt{CNN}(\cdot\,; \Theta)$ with network parameters Θ, and an observed noisy image $\boldsymbol{X}$ generated following the degradation model $\boldsymbol{X} = \boldsymbol{Z} + \boldsymbol{E}$, where $\boldsymbol{E}$ is i.i.d. Gaussian noise with a standard deviation σ. To learn the CNN denoiser $\texttt{CNN}(\cdot\,; \Theta)$, we assume that that the set of parameters Θ which allow it to recover $\boldsymbol{Z}$ from $\boldsymbol{X}$, are the same as those which allow $\texttt{CNN}(\cdot\,; \Theta)$ to recover $\boldsymbol{X}$ from $\boldsymbol{X} + \boldsymbol{E}$. This assumption has been used to learn image adapted CNNs for super resolution in [16]. It allow us to train the denoiser $\texttt{CNN}(\cdot\,; \Theta)$ using the image pair $(\boldsymbol{X} + \boldsymbol{E}, \boldsymbol{X})$ by minimizing the loss function:

$$\ell(\Theta) = \|\texttt{CNN}(\boldsymbol{X} + \boldsymbol{E}; \Theta) - \boldsymbol{X}\|_1. \tag{14}$$

We adopted the method in [17] to estimate σ in each channel of $\boldsymbol{X}$ to generate $\boldsymbol{E}$. The procedure for learning the proposed CNN-based denoiser is summarized in Algorithm 1.

Image-specific prior learning: Since there exist some inter-image variations between $\boldsymbol{Z}_h$ and $\boldsymbol{Z}_m$, we considered to train two independent denoisers, $\texttt{CNN}(\cdot\,; \Theta_h)$ and $\texttt{CNN}(\cdot\,; \Theta_m)$, to denoise $\boldsymbol{V}_h$ and $\boldsymbol{V}_m$, respectively. Considering that the equivalent noise levels of $\boldsymbol{V}_h$ and $\boldsymbol{V}_m$ decrease over the algorithm iterations, we propose to adaptively update the network parameters Θ_h and Θ_m to learn an image-specific prior at each iteration. This is performed by re-training $\texttt{CNN}(\cdot\,; \Theta_h)$ and $\texttt{CNN}(\cdot\,; \Theta_m)$ to denoise the estimates of the HRIs at the current iteration. To make the algorithm faster, we consider to train $\texttt{CNN}(\cdot\,; \Theta_h)$ and $\texttt{CNN}(\cdot\,; \Theta_m)$ in the first iteration and then fine-tune them in all the remaining itera-

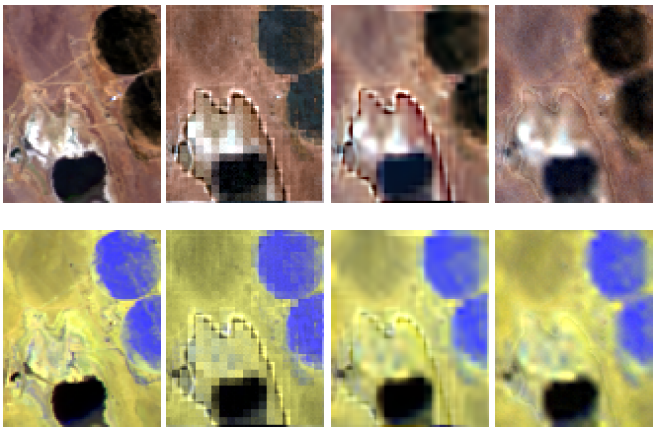

Fig. 1. Visible (top) and infrared (bottom) representation of the true and estimated HRI of the Lake Tahoe B scene.

Table 1. Quantitative comparisons of the competing methods.

Data set	Metrics	SAM	ERGAS	PSNR	UIQI
Ivanpah Playa	FuVar	1.6912	1.6820	26.2514	0.8219
	GSFus	2.0543	1.6761	26.2819	0.8606
	DIFIV	**1.3709**	**1.3085**	**28.4611**	**0.8942**
Lake Tahoe A	FuVar	7.8961	5.7982	20.1606	0.8072
	GSFus	6.2365	4.3429	22.6234	0.8958
	DIFIV	**5.6964**	**3.2543**	**25.2924**	**0.9435**
Lake Tahoe B	FuVar	5.0251	4.1696	20.9042	0.7421
	GSFus	3.7420	3.2103	23.2012	0.8394
	DIFIV	**2.7905**	**2.2007**	**26.6100**	**0.9167**

tions. The training strategy for the denoisers in Algorithm 1 is incorporated into the model-based optimization procedure, yielding the overall DIFIV strategy described in Algorithm 2.

4. EXPERIMENTS

We compared DIFIV to HMIF methods accounting for inter-image varibility, namely, FuVar [3] and GSFus [5], on three real data sets: the Ivanpah Playa and the Lake Tahoe A and B, described with more details in [4]. These data sets contained one reference HRI and an MI acquired by the AVIRIS and the Sentinel-2A instruments, respectively, with a spatial resolution of 20m [3]. The HI and MI contained $L_h = 173$ and $L_m = 10$ bands, respectively. For all acquired HRIs, which had the same spatial resolution as the MIs, a pre-processing procedure as described in [18] was performed. The observed HIs were generated according to (1), where $\boldsymbol{F}$ was an 8×8 Gaussian blurring operator with standard deviation 4 and $\boldsymbol{D}$ a downsampling operator with the scaling factor 4. The SRF $\boldsymbol{R}$ was acquired from calibration measurements of the Sentinel-2A instrument and known a priori. For all experiments, Gaussian noise was added to both HIs and MIs to obtain a signal-to-noise ratio (SNR) of 30 dB.

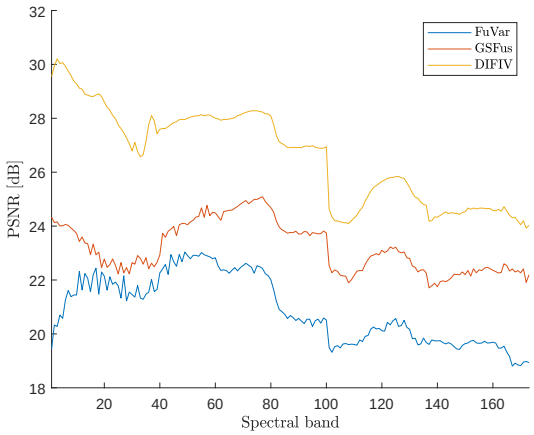

Fig. 2. PSNR curves as functions of the spectral bands for the Lake Tahoe B.

We implemented DIFIV 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 [19] with an initial learning rate 0.0002 was used to minimize the loss function in (14). 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 [20], trained for 10000 epochs in the first iteration and fine-tuned for 2000 epochs in the remaining iterations. We set $p = 1.5, \lambda = 0.01$ and $\lambda_m = \lambda_n = 0.1$ for the Ivanpah Playa. For the Lake Tahoe A and B, we set $p = 1.8, \lambda = 0.002$ and $\lambda_m = \lambda_n = 0.01$. For the other parameters, we set $\rho = 0.1$ and $\epsilon = 10^{-6}$.

The quantitative results (including the SAM, ERGAS, PSNR and UIQI metrics [3]) of the compared methods on all data sets are reported in Table 1. It can be seen that DIFIV achieves the best quantitative results, with considerable improvements observed in all three data sets. The visual inspection of results for the Lake Tahoe B data set is shown in Figure 1, where we observe that DIFIV provides spatial reconstructions closest to the ground truth and without significant artifacts, which are observed on the images reconstructed by both FuVar and GSFus. Figure 2 illustrates the PSNR curves per spectral bands over the Lake Tahoe B, from which it can be observed that the performance improvements obtained by DIFIV are consistent across all spectral bands.

5. 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. 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 using CNN-based denoising operations. The proposed method achieved superior experimental performance in the presence of inter-image variability when compared to state-of-the-art approaches.

6. 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 Geosci. Remote Sens. Mag.*, vol. 1, no. 2, pp. 6–36, 2013.

[2] N. Yokoya, C. Grohnfeldt, and J. Chanussot, "Hyperspectral and multispectral data fusion: A comparative review of the recent literature," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 2, pp. 29–56, 2017.

[3] R. A. Borsoi, T. Imbiriba, and J. C. M. Bermudez, "Super-resolution for hyperspectral and multispectral image fusion accounting for seasonal spectral variability," *IEEE Trans. Image. Process.*, vol. 29, no. 1, pp. 116–127, 2020.

[4] 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 J. Sel. Top. Sig. Process.*, vol. 15, no. 3, pp. 702–717, 2021.

[5] X. Fu, S. Jia, M. Xu, J. Zhou, and Q. Li, "Fusion of hyperspectral and multispectral images accounting for localized inter-image changes," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2021.

[6] Y. Romano, M. Elad, and P. Milanfar, "The little engine that could: Regularization by denoising (RED)," *SIAM J. Imaging Sci.*, vol. 10, no. 4, pp. 1804–1844, 2017.

[7] X. Wang, J. Chen, Q. Wei, and C. Richard, "Hyperspectral image super-resolution via deep prior regularization with parameter estimation," *IEEE Trans. Circuits Syst, Video Technol*, vol. 32, no. 4, pp. 1708–1723, 2021.

[8] 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 Geosci. Remote Sens. Mag.*, vol. 9, no. 4, pp. 223–270, 2021.

[9] S. Liu, D. Marinelli, L. Bruzzone, and F. Bovolo, "A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 140–158, 2019.

[10] Z. Lu, "Iterative reweighted minimization methods for ℓ_p regularized unconstrained nonlinear programming," *Math. Program.*, vol. 147, no. 1, pp. 277–307, 2014.

[11] 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.

[12] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Trans. Image. Process.*, vol. 26, no. 7, pp. 3142–3155, 2017.

[13] R. Imamura, T. Itasaka, and M. Okuda, "Zero-shot hyperspectral image denoising with separable image prior," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop. (ICCVW)*, 2019, pp. 0–0.

[14] 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.

[15] X. Wang, J. Chen, C. Richard, and D. Brie, "Learning spectral-spatial prior via 3ddncnn for hyperspectral image deconvolution," in *Proc. IEEE Int. Conf. on Acoust, Speech, Signal Process (ICASSP)*. IEEE, 2020, pp. 2403–2407.

[16] A. Shocher, N. Cohen, and M. Irani, ""zero-shot" super-resolution using deep internal learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 3118–3126.

[17] D. L. Donoho and J. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[18] M. Simões, J. Bioucas-Dias, L. B. Almeida, and J. Chanussot, "A convex formulation for hyperspectral image superresolution via subspace-based regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3373–3388, 2015.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1026–1034.