Deep Blind Hyperspectral Image Fusion

Wu Wang[1], Weihong Zeng[1], Yue Huang[1], Xinghao Ding[1]*, John Paisley[2]
[1]Fujian Key Laboratory of Sensing and Computing for Smart City,
School of Information Science and Engineering, Xiamen University, China
[2]Department of Electrical Engineering, Columbia University, New York, NY, USA
wangwu@stu.xmu.edu.cn, zengwh@stu.xmu.edu.cn
huangyue05@gmail.com, dxh@xmu.edu.cn, jpaisley@columbia.edu

Abstract

Hyperspectral image fusion (HIF) reconstructs high spatial resolution hyperspectral images from low spatial resolution hyperspectral images and high spatial resolution multispectral images. Previous works usually assume that the linear mapping between the point spread functions of the hyperspectral camera and the spectral response functions of the conventional camera is known. This is unrealistic in many scenarios. We propose a method for blind HIF problem based on deep learning, where the estimation of the observation model and fusion process are optimized iteratively and alternatingly during the super-resolution reconstruction. In addition, the proposed framework enforces simultaneous spatial and spectral accuracy. Using three public datasets, the experimental results demonstrate that the proposed algorithm outperforms existing blind and non-blind methods.

1. Introduction

Hyperspectral image (HSI) analysis has a wide range of applications for object classification and recognition [13, 9, 33, 17], segmentation [22], tracking [23, 24] and environmental monitoring [18] in both computer vision and remote sensing. While HSI facilitates these tasks through information across a large number of spectra, these many additional dimensions of information means that the potential spatial resolution of HSI systems is severely limited compared with RGB cameras. HIF addresses this challenge by using the jointly measured high resolution multispectral image (HR-MSI)—often simply RGB—to improve the low resolution HSI (LR-HSI) by approximating its high resolution version (HR-HSI).

Generally, most state-of-the art methods formulate the observation model through the linear functions [28, 7, 20]

(a) gt (b) HySure (c) Ours

Figure 1: The 31st band of a reconstructed high resolution hyperspectral image (HR-HSI) with unknown spectral response function. (a) ground-truth HR-HSI, (b) result of HySure [20], (c) our result.

$$\boldsymbol{Y} = \boldsymbol{XBS}, \tag{1}$$

$$\boldsymbol{Z} = \boldsymbol{RX}, \tag{2}$$

where $\boldsymbol{X}$ is the HR-HSI, $\boldsymbol{Y}$ is the LR-HSI and $\boldsymbol{Z}$ is the HR-MSI. The linear operators $\boldsymbol{B}$ and $\boldsymbol{S}$ perform the appropriate transformations to map $\boldsymbol{X}$ to the measured values; $\boldsymbol{B}$ represents a convolution between the point spread function of the sensor and the HR-HSI bands, $\boldsymbol{S}$ is a downsampling operation, and $\boldsymbol{R}$ is the spectral response function of the multispectral imaging sensor. The spectral response functions and point spread functions are often assumed to be at least partly known. A common way to learn $\boldsymbol{X}$ is through optimizing an objective function of the form

$$\min_{\boldsymbol{X}} \ \|\boldsymbol{Y} - \boldsymbol{XBS}\|_F^2 + \lambda_1 \|\boldsymbol{Z} - \boldsymbol{RX}\|_F^2 + \lambda_2 \varphi(\boldsymbol{X}), \tag{3}$$

where the first and second terms enforce agreement with the data and the third term is a regularization [12, 15, 6, 7]. However, this assumed relationship between $\boldsymbol{X}$,$\boldsymbol{Y}$,$\boldsymbol{Z}$ is not always true, and because the information available about the sensor is incomplete, it is unknowable [26]. In other words, this non-blind fusion is often only an approximation, and

*Corresponding author

therefore performance depends on approximation quality. Additionally, how to preserve both spectral and spatial information simultaneously is unresolved due to the trade-off between the two data fidelity terms.

Previous works usually assumed that the linear mapping between the point spread functions of the hyperspectral camera and the spectral response functions of the conventional camera is known, which is unrealistic in many scenarios. In this paper, we instead perform *blind* hyperspectral image fusion. We treat the problem as a low-level inverse problem with bias between the training and testing data. We address the problem by estimating the degradation process with additional regularization to improve model generalization. Compared with the latest blind and non-blind methods, experimental results on both simulated and real data demonstrate state-of-art performance and robustness. Although this is not a general inverse framework, with appropriate modifications the proposed work can benefit other low-level inverse problems with data bias, where the assumed degradation procedure is different from the true value.

2. Related Work

There have been numerous methods specifically designed for HSI super-resolution, including penalty based approaches [3, 20, 29, 28, 34], matrix factorization approaches [12, 15, 8, 14, 2], tensor factorization approaches [6, 16], and deep learning approaches [7]. Most relevant to our work is HySure [20], which attempts to estimates $\boldsymbol{B}$ and $\boldsymbol{R}$ from data via convex optimization based on two quadratic data-fitting terms and total variation regularization. To simplify the problem, HySure assumes that these two operators are linear. HySure also minimize an objective function similar to Eq. 3). Our model is based on an iterative back-projection refinement procedure similar to ideas used for other image processing problems. For example, [19] proposed a general iterative regularization framework for image denoising by iteratively refining a cost function. Recently, [21] proposed an iterative scheme for Reverse Filtering, which updates recovered images according to the filtering effect. In image super-resolution, iterative back-projection (IBP) refinement was proposed by [19]. Our approach is similarly motivated.

3. Motivation

Given a LR-HSI and HR-MSI image pair, the goal of the HSI fusion problem is to obtain an HR-HSI image $\boldsymbol{X} \in R^{(W\times H\times B)}$ that has both high spatial and high spectral resolution, with W, H and B the image width, image height and number of spectral bands, respectively. This can be formulated as

$$\hat{\boldsymbol{X}} = f(\boldsymbol{Y}, \boldsymbol{Z}), \tag{4}$$

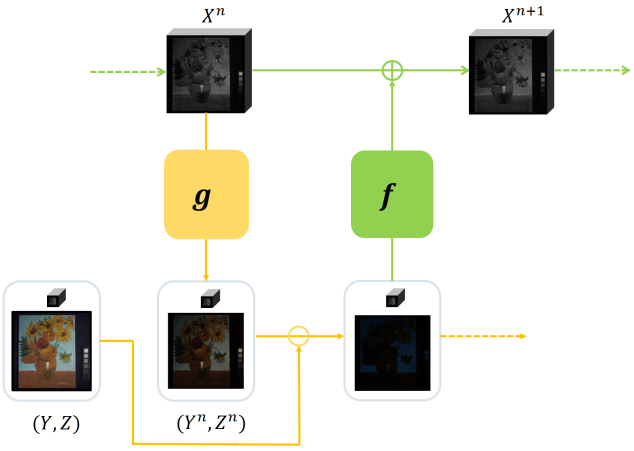


Figure 2: Illustration of our algorithm. This is the detailed operation process in one iteration

where $\boldsymbol{Y} \in R^{(w\times h\times B)}$ stands for the LR-HSI, $\boldsymbol{Z} \in R^{(W\times H\times b)}$ stands for the HR-MSI and $\hat{\boldsymbol{X}}$ stands for estimated HR-HSI. Generally HIF is highly under-constrained and difficult to solve because the total number of observations obtained from HR-MSI and LR-HSI is is much smaller than the unknowns ($whB + WHb \ll WHB$). For non-blind hyperspectral image fusion, given the parameters $\boldsymbol{B}$, $\boldsymbol{R}$, most methods learn the mapping function f in Eq. 4 by optimizing the objective function in Eq. 3. However, for the blind fusion of hyperspectral images, the parameters $\boldsymbol{B}$ and $\boldsymbol{R}$ are unknown, and so it is difficult to directly solve the objective function in Eq. 3. To address this problem, the observation model should also be learned. We formulate this process as

$$(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{Z}}) = g(\boldsymbol{X}), \tag{5}$$

where g stands for the observation model to be learned. For example, HySure first learns this mapping, estimating the parameters $\boldsymbol{B}$ and $\boldsymbol{R}$ of the observation model from the data. Then they introduce the learned parameters $\boldsymbol{B}$ and $\boldsymbol{R}$ into Eq. 3 to solve the forward fusion problem.

If the hyperspectral image blind fusion problem is perfectly solved, using $\hat{\boldsymbol{X}}$ obtained by the fusion function f in the function g, the resulting $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{Z}}$ should be the same as the inputs $\boldsymbol{Y}$ and $\boldsymbol{Z}$. But in practice there will be an error in the estimation of these two values. We thus propose an iterative fusion framework that iteratively learns these two functions by letting them correct each other. The proposed framework is formulated as

$$\boldsymbol{X^{n+1}} = \boldsymbol{X^{n}} + f(\Delta \boldsymbol{Y}^n, \Delta \boldsymbol{Z}^n), \tag{6}$$

where $\boldsymbol{X}^n$ stands for the HR-HSI in the n^{th} iteration, and $\Delta \boldsymbol{Y}^n$ and $\Delta \boldsymbol{Z}^n$ stand for the back-projection error. The cal-

culation of $\Delta \boldsymbol{Y}^n$ and $\Delta \boldsymbol{Z}^n$ can be written as

$$\begin{aligned}(\Delta \boldsymbol{Y}^n, \Delta \boldsymbol{Z}^n) &= (\boldsymbol{Y}, \boldsymbol{Z}) - (\boldsymbol{Y}^n, \boldsymbol{Z}^n) \\ &= (\boldsymbol{Y}, \boldsymbol{Z}) - g(\boldsymbol{X}^n),\end{aligned} \tag{7}$$

where $\boldsymbol{Y}^n$ and $\boldsymbol{Z}^n$ represent the learned LR-HSI and HR-MSI in the n^{th} iteration. Therefore, we both learn the fusion function f and the observation model g as in other blind methods. However, the recovered image $\hat{\boldsymbol{X}}$ will still suffer from spatial and spectral distortion. We therefore iteratively correct the result of the fusion process done by the observation model. To describe our algorithm, we use the illustration in Fig. 2. We start from HR-HSI $\boldsymbol{X}^n$, which is the fused image in the n^{th} iteration. After applying the (unknown) back-projection function g to $\boldsymbol{X}^n$, we obtain $(\boldsymbol{Y^n}, \boldsymbol{Z^n})$. We then calculate the residual $(\boldsymbol{Y}, \boldsymbol{Z}) - (\boldsymbol{Y^n}, \boldsymbol{Z^n})$, which contains both spectral and structural distortion. Finally, we fuse the residual with the (unknown) HSI fusion function f, and add this to correct $\boldsymbol{X}^n$. Then we perform another iteration with similar steps. Empirically, $\boldsymbol{X}^n$ with increasing n better approximates the ground truth $\boldsymbol{X}$.

4. Deep Blind Iterative Fusion Network (DBIN)

Using the above algorithm, we create a deep neural network for HSI fusion by unfolding all steps of the algorithm as network layers. The proposed network is a structure of n stages implemented in n iterations using Eq. (6). The reason we chose the convolutional neural network to implement this framework is twofold. First, while the objective function of most methods contains two data fidelity terms that must trade off between spectral and structural quality, for convolutional neural networks it is easy to construct an objective function that contains only one data fidelity term, which avoids this trade-off. Second, matrix factorization-based methods cannot fully exploit the spatial-spectral correlation of the HSIs since they need to unfold the three-dimensional HSI into matrices, while convolutional neural networks are very suitable for extracting spatial-spectral correlations. Furthermore, deep learning is much faster to optimize than traditional iterative algorithms in this area.

The pipeline of the proposed model is illustrated in Fig. 3 (top). The model takes initialized HR-HSI $\boldsymbol{X}^0$ as input and refines this initialized value with the "iterative refinement unit" (IRU) according to

$$\boldsymbol{X^{n+1}} = \boldsymbol{X^n} + f_\theta(\boldsymbol{Y}, \boldsymbol{Z}, g_\theta(\boldsymbol{X}^n)), \tag{8}$$

where θ denotes the trainable parameter set of the CNN. The initialized HR-HSI $\boldsymbol{X}^0$ was learned together with the IRU. This can be formulate as

$$\boldsymbol{X^0} = f_\theta(\boldsymbol{Y}, \boldsymbol{Z}), \tag{9}$$

In order to simulate the iterative optimization process shown in Eq. (6), all parameters of the IRU are shared. Finally, we combine the fused images HR-HSI $\boldsymbol{X}^n$ produced in each of the intermediate stages to obtain the final result, which we call *dense fusion*. The core of our network is the iterative refinement unit. Next we discuss the IRU and dense fusion mechanisms in more detail.

4.1. Iterative Refinement Unit (IRU)

To make up for information loss during fusion, each stage the IRU takes the output of the previous stage, $\boldsymbol{X}^n$, plus the HR-MSI image $\boldsymbol{Z}$ and LR-HSI image $\boldsymbol{Y}$ as input to obtain an refined $\boldsymbol{X}^{n+1}$. This output becomes one input of the next layer according to Eq. (8). As the number of iterations increase, spectral and structural distortion reduces. We show the detailed structure of the IRU in Fig. 3 (bottom). The IRU consists of a measuring module and a fusion module. The measurement module is responsible for learning the observation model, while the fusion module extracts useful spatial and spectra information.

Measurement module The observation model has previously been used as a constraint [35, 11]. In the proposed work, we apply a similar idea to constrain the blind fusion of hyperspectral image. Referring to Eq. (5), this process can be written as

$$(\boldsymbol{Y^n}, \boldsymbol{Z^n}) = g_\theta(\boldsymbol{X^n}), \tag{10}$$

where θ denotes the trainable network parameter. For Eq. (1) in the observation model, many algorithms assume that $\boldsymbol{B}$ is a convolution operator and $\boldsymbol{S}$ is a down-sampling operator. Similarly, we model this process with a single-layer convolution with stride. For Eq. (2), many algorithms treat $\boldsymbol{R}$ as a matrix. Since our model is based on the convolutional neural network, which is particularly good at processing 3D tensors, we also use single layer convolution models Eq. (2) without resorting to matrix representations. In addition to using convolution to simulate the three parameters $\boldsymbol{B}$, $\boldsymbol{S}$ and $\boldsymbol{R}$, we also apply nonlinear activation functions after the convolution based on our nonlinear assumptions of the observation model.

Fusion Module We use this module to extract spatial structure and spectral information from the residuals to refine the previous results. Following Eq. (6), this process can be written as

$$\boldsymbol{X^{n+1}} = \boldsymbol{X^n} + f_\theta(\Delta \boldsymbol{Y}^n, \Delta \boldsymbol{Z}^n). \tag{11}$$

The fusion module is built using ResNet [10] since our network is very deep and may suffer from gradient vanishing during training. We first concatenate the upsampled residuals of the LR-HSI with the residuals of the multi-spectral

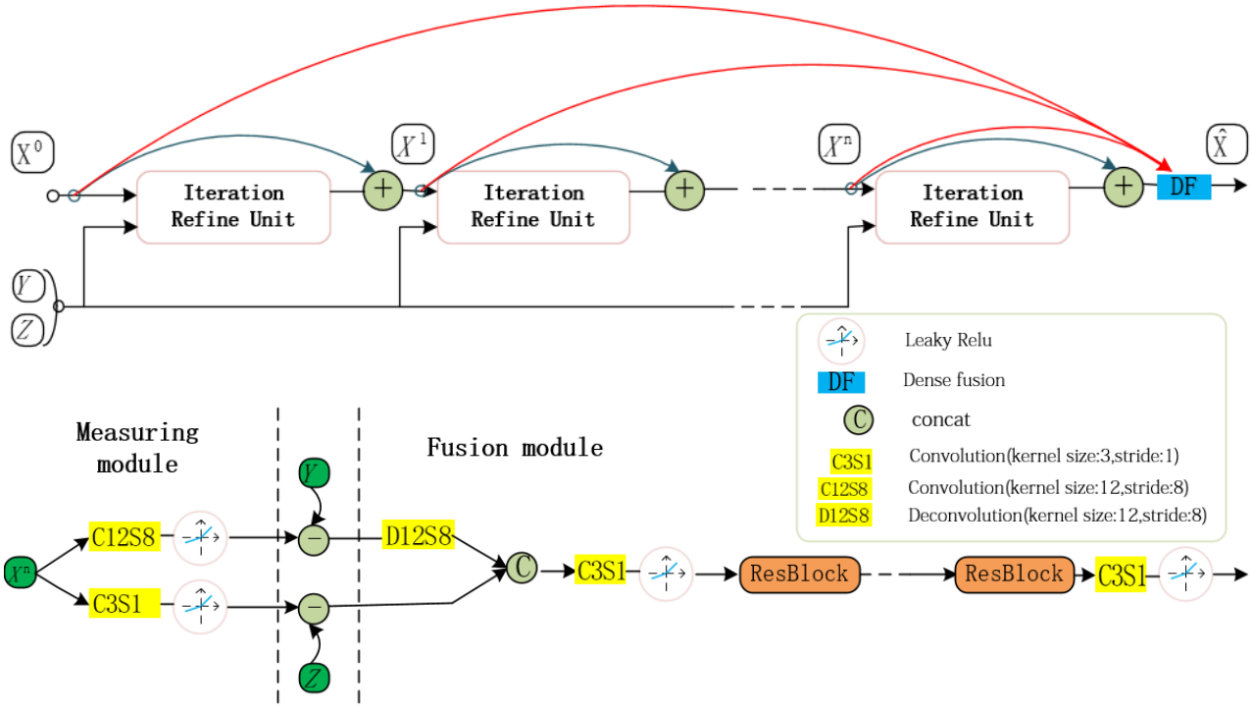


Figure 3: Detailed structure of the proposed network. (top) Overall network structure. (bottom) Detailed structure of the iterative refinement unit (IRU).

image. Then we apply several ResBlocks to extract features. These features will be used to refine the HR-HSI output at the last iteration.

4.2. Dense Fusion

When training the network we fuse the outputs of each iteration phase using convolution for the final output. We call this mechanism "dense fusion." This is illustrated in Fig. 3. This is motivated by the fact that deep neural networks extract features having different information at different depths. Similarly, in our network hyperspectral images generated in different iterations may have different spatial or spectral information, and the results can be combined to further improve the performance of the network. We will verify the effectiveness of deep fusion in the experiments section.

Therefore, the final generated HR-HSI can be written as

$$\hat{\boldsymbol{X}} = Conv(Concat(\boldsymbol{X^0}, X^1, ..., \boldsymbol{X^n}, ...)), \quad (12)$$

where "Conv" represents the convolution operator and "Concat" represents the concatenation operator. We use pixel-wise L_1 reconstruction loss for $\hat{\boldsymbol{X}}$ during training. The L_1 loss can better preserve the edges of an image, which is desired in our task. The overall loss function is defined as follows:

$$\mathcal{L} = \|\hat{\boldsymbol{X}} - X\|_1, \quad (13)$$

where $\boldsymbol{X}$ are ground truth HR-HSI, while $\hat{\boldsymbol{X}}$ is the corresponding output HR-HSI.

5. Experiments

5.1. Data and Experimental Setup

We use three publicly available hyperspectral databases for our simulation experiments: CAVE [30] ,Harvard [30], and NTIRE2018 [1]. For real data experiments, we use WV2.[1]

The Harvard database contains 50 indoor and outdoor images recorded under daylight illumination, and 27 images under artificial or mixed illumination. We only use the 50 indoor images for our experiments. We use the first 30 HSI for training, and the last 20 HSI for testing.

The CAVE database includes 32 indoor images captured under controlled illumination. We use the first 20 HSI for training and the last 12 HSI for testing. The CAVE dataset is generally considered to be more challenging than the Harvard dataset because the Harvard images have higher spatial resolution, while pixels in close range usually have similar spectral reflectance and therefore typically contain smoother reflections.

The NTIRE2018 database was built for the NTIRE2018

[1] https://www.harrisgeospatial.com/Data-Imagery/Satellite-Imagery/High-Resolution/WorldView-2

challenge on spectral reconstruction from RGB images. This dataset contains two parts: "Train1" includes 201 images from the ICVL dataset [4], which consists of RGB images created by applying a known spectral response function to ground truth hyperspectral images. "Train2" has 53 RGB images created by applying an unknown response function to ground truth hyperspectral information. We use Train1 for training, and Train2 for testing.

The WV2 database contains an 8-band LR-MSI and RGB image pair. We use the upper part as the training set and the lower part as the test set. Since the ground truth HR-MSI is not available in the real dataset, we use Wald's protocol [32] to generate the training data.

We use several non-blind state-of-art methods for comparison: sparse fusion (SPARTF) [28], coupled sparse tensor factorization (CSTF) [16], coupled spectral unmixing (CSU) [15], nonnegative-structured sparse representation (NSSR) [8], and DHSIS [7]. We also compare with HySure [20] and DTV [5], which are blind HIF methods. For quantitative comparison, PSNR, structural similarity index (SSIM [27]), spectral angle mapper index (SAM [31]) and *erreur relative globale adimensionnelle de synth`ese* (ERGAS [25]) are used for evaluation. SAM is a spectral evaluation method used in remote sensing, which measures the angular difference between the estimated image and the ground truth [31]. SSIM is an indicator of the spatial structures preservation of the estimated image. ERGAS reflects the overall quality of the fused image.

5.2. Model verification with CAVE dataset

We first conduct simulated experiments to verify our deep blind iterative fusion network (DBIN) quantitatively. We compare the performance of the proposed DBIN with different stages number n. We also denote our model with dense fusion "DBIN+."

Table 1 shows the average results over 12 testing HSI images. We observe that DBIN with more stages has better performance, while the parameter size in the network has not increased. This shows that our network can indeed iteratively refine the target. We further observe that dense fusion can significantly improve the model. In the following experiments, we will use "DBIN+" to compare with other methods.

5.2.1 Can our model learn the observation model?

To investigate this question, we visualize low-resolution multispectral images generated from different iterations in Fig. 4. We only show the residuals of the first three iterations because the subsequent residuals are small enough to be ignored. It can be seen from the residual image of the first iteration that the reconstruction error is large, indicating that the HR-HSI learned by the network has very serious

Table 1: Model analysis on the CAVE dataset. n represents the number of iterations of the network. "+" indicates that the network has a dense fusion structure.

Method	PSNR	SSIM	SAM	ERGAS
Best Values	$+\infty$	1	0	0
DBIN (n=1)	45.58	0.9927	3.55	0.74
DBIN (n=3)	45.69	0.9925	3.55	0.69
DBIN (n=5)	46.32	0.9930	3.41	0.66
DBIN+ (n=5)	47.51	0.9934	3.18	0.58

spatial and spectral distortion. The residual images in the second iteration are small, and only contain a small amount of spectral information, indicating that the network has been able to preserve this information, but there is still spatial information distortion. At the third iteration, the residuals are already small and the network has learned the observation model. In fact, we should use multiple convolution kernels of different sizes to learn the observation model as we do not know its size, but we found in our experiments that adopting this strategy did not lead to performance gain. Therefore, for all following experiments we use single layer 12×12 convolution with stride 8 to simulate $\boldsymbol{B}, \boldsymbol{S}$, and another single layer 3×3 convolution with stride 1 to learn $\boldsymbol{R}$.

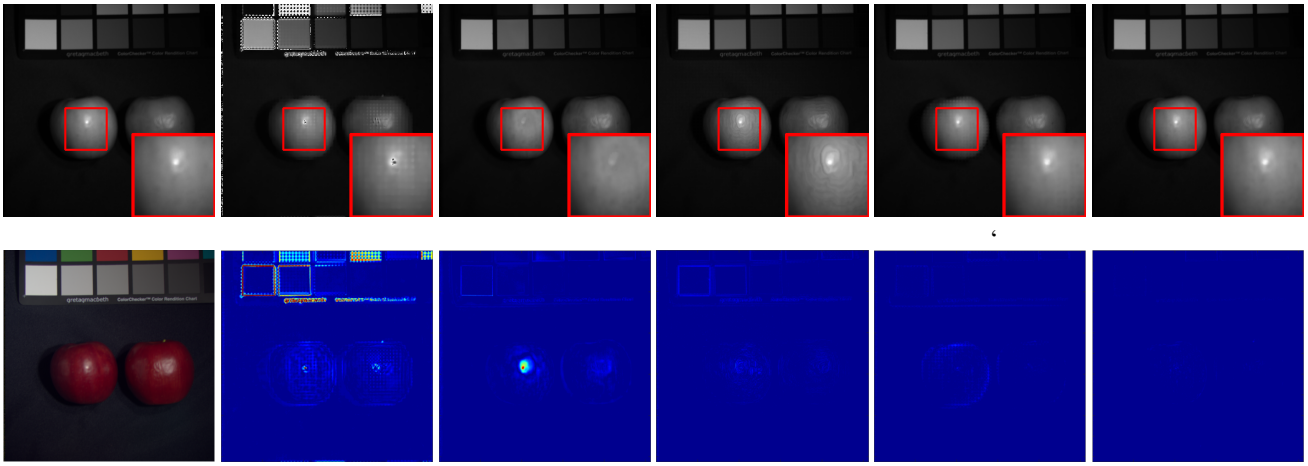
(a) Z (b) Z^1 (c) Z^2 (d) Z^3

(e) $Z - Z^1$ (f) $Z - Z^2$ (g) $Z - Z^3$

Figure 4: HR-MSI learned by the network in different iterations. (a) The ground truth HR-MSI. (b)-(c) Learned HR-MSI at n^{th} (n=1,2,3) iteration. (d)-(f) The residual corresponding to (a)-(c). For better visual quality, we have magnified these residual by a factor of three.

5.3. Non-blind fusion on CAVE and Harvard data

We follow the same setting as [7]. First we apply an 8×8 Gaussian filter with a mean of 0 and a standard deviation of 2, and then downsample every 8 pixels in both the vertical and horizontal directions for each band of the reference to simulate the LR-HSI. The RGB images $\boldsymbol{Z}$ were simulated



(a) gt and RGB (b) HySure (c) CSU (d) NSSR (e) DHSIS (f) Ours

Figure 5: Qualitative results of the CAVE dataset at band 31. Top row: reconstructed images. Bottom row: reconstruction errors – light color indicates less error, dark color indicates larger error.

Table 2: Average quantitative results of various methods on the CAVE test dataset.

	Method	PSNR	SSIM	SAM	ERGAS
Non-blind methods	SPARTF	39.51	0.946	10.2	1.28
	CSTF	42.66	0.971	6.68	0.98
	CSU	41.86	0.982	6.30	1.14
	NSSR	43.82	0.987	4.07	0.84
	DHSIS	45.59	0.990	3.91	0.73
Blind methods	HySure	37.35	0.945	9.84	2.01
	DBIN+	47.51	0.993	3.18	0.58

Table 3: Average quantitative results of various methods on the Harvard test dataset.

	Method	PSNR	SSIM	SAM	ERGAS
Non-blind methods	SPARTF	41.08	0.943	5.29	2.93
	CSTF	40.10	0.942	4.92	3.08
	CSU	45.10	0.981	3.68	1.40
	NSSR	46.31	0.982	3.46	1.20
	DHSIS	46.02	0.981	3.54	1.17
Blind methods	HySure	43.88	0.975	4.20	1.56
	DBIN+	46.67	0.983	3.42	1.15

by integrating over the original spectral channels using the spectral response $\boldsymbol{R}$ of a Nikon D700 camera.[2] For this setting, the parameters $\boldsymbol{B}$ and $\boldsymbol{R}$ are known.

The average quantitative values across the two datasets are shown in Table 2 and Table 3. The experimental results demonstrate that the proposed approach achieves significantly better results than other methods on the CAVE dataset according to all index measures, suggesting that our method can better preserve both spatial and spectral information. As the Harvard dataset is less challenging than the CAVE dataset, all the compared methods achieve good results, but the proposed algorithm still performs better. This demonstrates that DBIN+ can handle challenging scenarios much better than state-of-the-art. (Actually these experiments are "unfair" for our method, since we do not use the knowledge of $\boldsymbol{B}$, $\boldsymbol{R}$ unlike these other methods.)

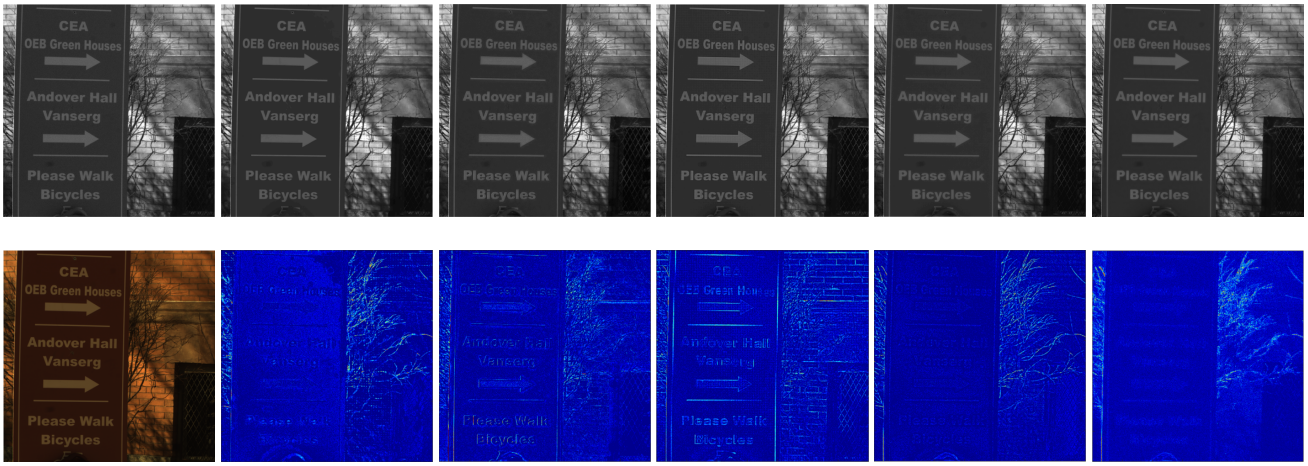
We also show qualitative results of both datasets in Fig. 5 and Fig. 6. (Since SPARTF and CSTF perform worse that other non-blind methods, we do not provide the qualitative results of these two methods.) Both of the output images of HySure and DHSIS suffer from grid-like structural distortion, and NSSR and CSU have ring shaped distortion. Meanwhile, our results are almost identical in visual quality to the ground-truth images. We also achieve minimal reconstruction error at both the edges and smooth areas of the image. This indicates that our algorithm has less structural and spectral distortion than other methods.

5.4. Semi-blind fusion on NTIRE2018 data

As with the previous experiments we need to simulate a LR-HSI, but we use the RGB image of this dataset directly for training and testing. For this set of experiments, the parameter $\boldsymbol{B}$ is known and the parameter $\boldsymbol{R}$ is unknown. We thus call this experiment semi-blind fusion. For the non-blind methods with which we compare, we directly use the $\boldsymbol{R}$ matrix built into their code to test.

Table 4 shows the average performance over 53 test images of the competing methods. We observe that the proposed method significantly outperforms other methods with

[2] http://www.maxmax.com/spectral_response.htm



(a) gt and RGB (b) HySure (c) CSU (d) NSSR (e) DHSIS (f) Ours

Figure 6: Qualitative results of the Harvard dataset at band 31. Top row: reconstructed images. Bottom row: reconstruction errors – light color indicates less error, dark color indicates larger error.

Table 4: Average quantitative results of the test methods on the NTIRE 2018 dataset.

	Method	PSNR	SSIM	SAM	ERGAS
Non-blind methods	SPARTF	23.6	0.59	7.18	6.94
	CSTF	9.41	0.24	15.4	29.2
	CSU	19.7	0.76	9.36	8.19
	NSSR	22.9	0.41	17.4	9.78
	DHSIS	1.13	0.19	20.4	257
Blind methods	HySure	37.3	0.94	5.14	2.13
	DBIN+	46.4	0.98	2.41	0.71

respect to all evaluation measures by a great margin. While those non-blind methods use a predefined $\boldsymbol{R}$ matrix, thus achieving better results on the non-blind experiments, they generate worse results on this experiment. In fact, DHSIS do not work at all in this set of experiments since they first use the preset $\boldsymbol{R}$ matrix to solve the optimization problem for the initial value. Since the $\boldsymbol{R}$ matrix is inaccurate, the initial results obtained are not very good. When they then use neural networks to optimize this initial result the network does not converge, leading to worse performance. Hysure, while performing worse than the CSU and NSSR without knowing the exact $\boldsymbol{R}$ and $\boldsymbol{B}$, still achieved similar performance to the previous two sets of experiments when the $\boldsymbol{R}$ is unknown.

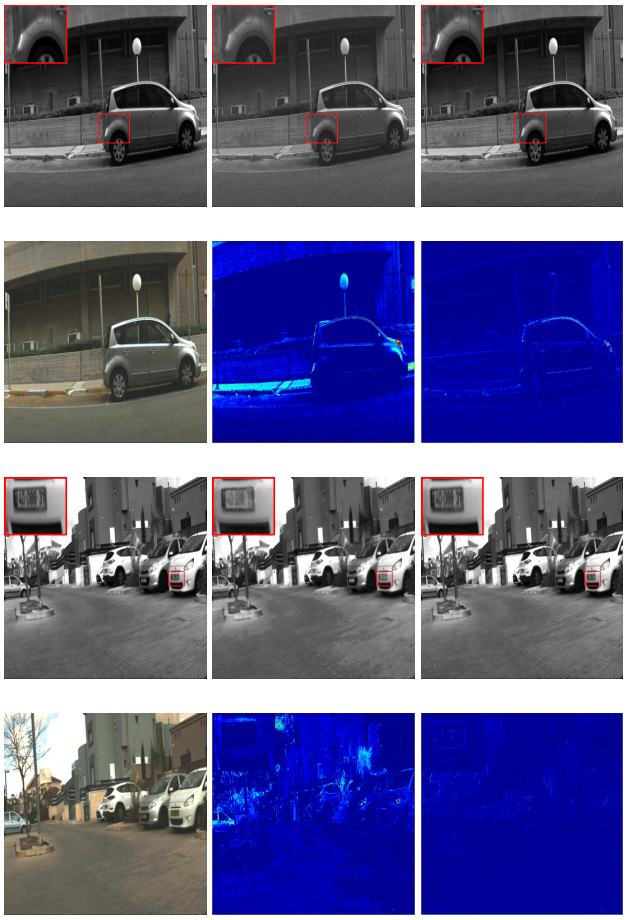
We also show the images of two test samples obtained by our method and HySure (band 31) in Fig 7. It is seen that the image obtained by DBIN+ is closest to the ground-truth, while the results of HySure usually contain obvious incorrect structure and suffer from spectral distortion. This is due to various reasons. First, although HySure attempts to estimate the observation model from the data, they do the estimation only once, but our model learns the observation model and the fusion process through an iterative alternating manner, allowing the alternative optimization between the two processes so the results of both estimation and fusion are more accurate. Second, the linear assumptions of HySure about the observation model may be limiting in real-world scenarios. Third, the two data fitting terms cause HySure to make trade-off between spectral and spatial preservation.

5.5. Real blind fusion on WV2 data

Here we provide the results on a public dataset of real multispectral images called WV2. Multispectral image fusion (MIF) aims to fuse a RGB (or PAN) image with an LR-MSI image to reconstruct a HR-MSI image. The slight difference between MIF and HIF is that hyperspectral images have many more bands than multispectral images. Other non-blind methods require a degradation matrix which is unknown in this case, thus we only provide the comparisons with blind methods. DTV [5] is the state-of-art blind method, so we only report the result of DTV for real experiments (see Fig. 8) since it runs very slowly and takes a few days to fuse a single image. Experiments demonstrate that HySure suffers from grid distortion, and DTV produces over-smooth effect while our method achieves the most satisfactory result.

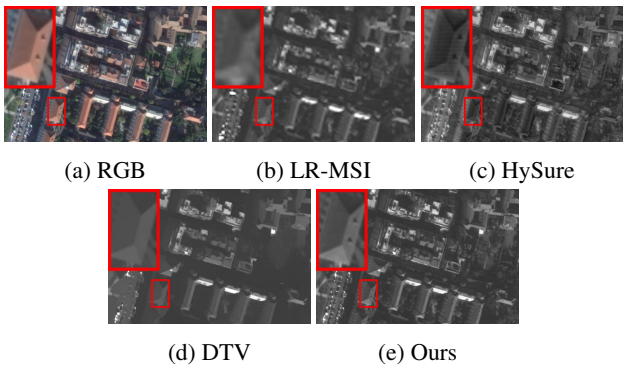
5.6. Extension: Single Image Super Resolution

Finally, our network can be directly extended to other ill-posed inverse problems. Fig. 9 shows the experimental results of single image super-resolution. Compared with MDSR (Winner of NTIRE2017 Super-Resolution Chal-



(a) gt and RGB (b) HySure (c) Ours

Figure 7: Qualitative results of the NTIRE 2018 dataset at band 31. Top: reconstructed images. Bottom: reconstruction errors – light color indicates less error, dark color indicates larger error.



(a) RGB (b) LR-MSI (c) HySure

(d) DTV (e) Ours

Figure 8: (a) and (b) are the real RGB and LR-MSI image acquired by World View-2. (c)-(e) The fused HR-MSI image. We only show band 3 of the MSI for simplicity

(a) LR (b) MDSR (c) Ours

Figure 9: Experimental results for $\times 4$ single image super resolution. (a) The bicubic downsampling low resolution image.(b) The result obtained by MDSR [20]. (c) The result of our method, DBIN+.

lenge) [20], our model achieves similar visual effects, despite not being specifically designed for this task. This demonstrates that the proposed network is more general and may by applied to other image processing tasks.

5.7. Conclusion

In this work, we proposed an iterative fusion framework for blind hyperspectral image fusion. We are able to iteratively and alternatingly estimate the observation model and predict the fusion model. We apply deep neural networks in this framework and design the entire iterative procedure as an end-to-end system. The proposed DBIN+ blindly fuses the LR-HSI with the HR-MSI without any prior knowledge about the observation model and preserves spectral and spatial information at the same time. Evaluations on four public datasets demonstrate that the proposed model achieves state-of-the-art performance in terms of quantitative result and visual quality.

6. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 61571382, 81671766, 61571005, 81671674, 61671309 and U1605252, in part by the Fundamental Research Funds for the Central Universities under Grants 20720160075 and 20720180059, in part by the CCF-Tencent open fund, and the Natural Science Foundation of Fujian Province of China (No. 2017J01126).

References

[1] NTIRE2018 challenge on spectral reconstruction from rgb images. `http://www.vision.ee.ethz.ch/ntire18/`.

[2] Naveed Akhtar, Faisal Shafait, and Ajmal Mian. Sparse spatio-spectral representation for hyperspectral image super-resolution. In *European Conference on Computer Vision*, pages 63–78, 2014.

[3] Naveed Akhtar, Faisal Shafait, and Ajmal Mian. Bayesian sparse representation for hyperspectral image super resolu-

tion. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3631–3640, 2015.

[4] Boaz Arad and Ohad Ben-Shahar. Sparse recovery of hyperspectral signal from natural rgb images. In *European Conference on Computer Vision*, pages 19–34, 2016.

[5] Leon Bungert, David A Coomes, Matthias J Ehrhardt, Jennifer Rasch, Rafael Reisenhofer, and Carola-Bibiane Schnlieb. Blind image fusion for hyperspectral imaging with the directional total variation. *Inverse Problems*, 2018.

[6] Renwei Dian, Leyuan Fang, and Shutao Li. Hyperspectral image super-resolution via non-local sparse tensor factorization. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3862–3871, 2017.

[7] Renwei Dian, Shutao Li, Anjing Guo, and Leyuan Fang. Deep hyperspectral image sharpening. *IEEE Transactions on Neural Networks and Learning Systems*, 29(11):5345–5355, 2018.

[8] Weisheng Dong, Fazuo Fu, and Xin Li. Hyperspectral image super-resolution via non-negative structured sparse representation. *IEEE Transactions on Image Processing*, 25(5):2337–2352, 2016.

[9] Mathieu Fauvel, Yuliya Tarabalka, Jon Atli Benediktsson, and James Tilton. Advances in spectral-spatial classification of hyperspectral images. *Proceedings of the IEEE*, 101(3):652–675, 2013.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE International Conference on Computer Vision*, pages 770–778, 2016.

[11] Qi-Xing Huang and Leonidas Guibas. Consistent shape maps via semidefinite programming. *Computer Graphics Forum*, 32(5):177–186, 2013.

[12] Rei Kawakami, John Wright, Yu-Wing Tai, Yasuyuki Matsushita, Moshe Ben-Ezra, and Katsushi Ikeuchi. High-resolution hyperspectral imaging via matrix factorization. In *CVPR 2011*, pages 2329–2336, 2011.

[13] Chiman Kwan, Bulent Ayhan, Jing Wang, and Chein-I Chang. A novel approach for spectral unmixing classification and concentration estimation of chemical and biological agents. *IEEE Transactions on Geoscience and Remote Sensing*, 44(2):409–419, 2006.

[14] Hyeokhyen Kwon and Yu-Wing Tai. Rgb-guided hyperspectral image upsampling. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 307–315, 2015.

[15] Charis Lanaras, Emmanuel Baltsavias, and Konrad Schindler. Hyperspectral super-resolution by coupled spectral unmixing. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3586–3594, 2011.

[16] Shutao Li, Renwei Dian, Leyuan Fang, and Jos M. Bioucas-Dias. Fusing hyperspectral and multispectral images via coupled sparse tensor factorization. *IEEE Transactions on Image Processing*, 27(8):4118–4130, 2018.

[17] Emmanuel Maggiori, Guillaume Charpiat, Yuliya Tarabalka, and Pierre Alliez. Recurrent neural networks to correct satellite image classification maps. *IEEE Transactions on Geoscience and Remote Sensing*, 55(9):4962–4971, 2017.

[18] Antonio Plaza, Qian Du, Jos M. Bioucas-Dias, Xiuping Jia, and Fred A. Kruse. Foreword to the special issue on spectral unmixing of remotely sensed data. *IEEE Transactions on Geoscience & Remote Sensing*, 49(11):4103–4110, 2011.

[19] Yaniv Romano and Michael Elad. Boosting of image denoising algorithms. *Siam Journal on Imaging Sciences*, 8(2):1187–1219, 2015.

[20] Miguel Simoes, Jose Bioucas-Dias, Luis B. Almeida, and Jocelyn Chanussot. A convex formulation for hyperspectral image superresolution via subspace-based regularization. *IEEE Transactions on Geoscience & Remote Sensing*, 53(6):3373–3388, 2015.

[21] Xin Tao, Chao Zhou, Shen Xiaoyong, Wang Jue, and Jia Jiaya. Zero-order reverse filtering. In *IEEE International Conference on Computer Vision*, 2017.

[22] Yuliya Tarabalka, Jocelyn Chanussot, and Jn Atli Benediktsson. Segmentation and classification of hyperspectral images using minimum spanning forest grown from automatically selected markers. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(5):1267–1279, 2010.

[23] Burak Uzkent, Matthew J. Hoffman, and Anthony Vodacek. Real-time vehicle tracking in aerial video using hyperspectral features. In *Computer Vision and Pattern Recognition Workshops*, pages 1443–1451, 2016.

[24] Burak Uzkent, Aneesh Rangnekar, and M. J. Hoffman. Aerial vehicle tracking by adaptive fusion of hyperspectral likelihood maps. In *Computer Vision and Pattern Recognition Workshops*, pages 233–242, 2017.

[25] Lucien Wald. Quality of high resolution synthesised images: Is there a simple criterion ? In *Fusion of Earth Data: Merging Point Measurements, Raster Maps and Remotely Sensed Images*, pages 99–103, 2009.

[26] Tianxing Wang, Guangjian Yan, Huazhong Ren, and Xihan Mu. Improved methods for spectral calibration of on-orbit imaging spectrometers. *IEEE Transactions on Geoscience and Remote Sensing*, 48(11):3924–3931, 2010.

[27] Zhou Wang, Alan Conrad Bovik, and Hamid Rahim Sheikh andEero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[28] Qi Wei, Jose Bioucas-Dias, Nicolas Dobigeon, and Jean-Yves Tourneret. Hyperspectral and multispectral image fusion based on a sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 53(7):3658–3668, 2015.

[29] Qi Wei, Nicolas Dobigeon, and Nicolas Dobigeon. Bayesian fusion of multi-band images. *IEEE Journal of Selected Topics in Signal Processing*, 9(6):1117–1127, 2015.

[30] Fumihito Yasuma, Tomoo Mitsunaga, Daisuke Iso, and Shree K. Nayar. Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum. *IEEE Transactions on Image Processing*, 19(9):2241–2253, 2010.

[31] Roberta H. Yuhas, Alexander F.H. Goetz, and Joe W. Boardman. Descrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In *Summaries of the Third Annual JPL Airborne Geoscience Workshop*, pages 147–149, 1992.

[32] Yongnian Zeng, Wei Huang, Maoguo Liu, Honghui Zhang, and Bin Zou. Fusion of satellite images in urban area: Assessing the quality of resulting images. In *International Conference on Geoinformatics*, 2010.

[33] Fan Zhang, Bo Du, and Liangpei Zhang. Scene classification via a gradient boosting random convolutional network framework. *IEEE Transactions on Geoscience and Remote Sensing*, 54(3):1793–1802, 2016.

[34] Yifan Zhang, Steve De Backer, and Paul Scheunders. Noise-resistant wavelet-based bayesian fusion of multispectral and hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 47(11):3834–3843, 2009.

[35] Jun-Yan Zhu*, Taesung Park*, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.