SSF-CNN: SPATIAL AND SPECTRAL FUSION WITH CNN FOR HYPERSPECTRAL IMAGE SUPER-RESOLUTION

Xian-Hua Han[1] *and Boxin Shi*[2] *and YinQiang Zheng*[3]

[1]Graduate School of Science and Technology for Innovation, Yamaguchi University, Japan
[2]Institute of Digital Media, School of EECS, Peking University, China
[3]National Institute of Informatics, Tokyo, Japan

ABSTRACT

Fusing a low-resolution hyperspectral image with the corresponding high-resolution RGB image to obtain a high-resolution hyperspectral image is usually solved as an optimization problem with prior-knowledge such as sparsity representation and spectral physical properties as constraints, which have limited applicability. Deep convolutional neural network extracts more comprehensive features and is proved to be effective in upsampling RGB images. However, directly applying CNNs to upsample either the spatial or spectral dimension alone may not produce pleasing results due to the neglect of complementary information from both low resolution hyper spectral and high resolution RGB images. This paper proposes two types of novel CNN architectures to take advantages of spatial and spectral fusion for hyperspectral image superresolution. Experiment results on benchmark datasets validate that the proposed spatial and spectral fusion CNNs outperforms the state-of-the-art methods and baseline CNN architectures in both quantitative values and visual qualities

Index Terms— Spatial and spectral fusion CNN, super resolution, hyperspectral image, SSF-CNN

1. INTRODUCTION

Hyperspectral (HS) imaging is an emerging technique that simultaneously obtains a set of images of the same scene on a large number of narrow band wavelengths. The acquired dense spectral bands of data significantly enrich the captured scene information and greatly enhance performance in many computer vision tasks such as object recognition and classification [1, 2, 3], tracking [4], segmentation [5], medical image analysis [6], and remote sensing [7, 8]. While HS imaging achieves rich spectral information, for guaranteeing sufficiently high signal-to-noise ratio, photon collection in hyperspectral sensors is usually performed in a much larger spatial region and thus results in much lower spatial resolution than RGB or Multi-Spectral (MS) images. Such low spatial resolution provided by existing sensors restricts application of existing computer vision algorithms for scene analysis and understanding. Fortunately, multi-spectral (e.g., RGB and RGB-NIR) images can be easily captured in high spatial resolution together with a hyperspectral imaging system. Therefore, fusing the low-resolution hyperspectral (LR-HS) and high-resolution multi-spectral (HR-MS) images to generate a high-resolution hyperspectral (HR-HS) image, which is called hyperspectral image super-resolution (HSI SR), is a popular solution to achieve both high spatial and spectral information.

Recent HSI SR methods are optimization based. Motivated by spectral decomposition with different constraints such as sparsity representation [9, 10, 11], spectral physical properties [9], spatial context similarity [11], the reconstruction errors of the spectral representation for both LR-HS and HR-MS (or HR-RGB) images [12, 11, 13] are jointly minimized. The quality of the recovered HR-HS image by optimization based methods greatly depends on the pre-defined constraints. Furthermore, the optimization procedure usually involves high computational cost due to the large number of constraint terms.

Recently, deep convolutional neural network (DCNN) has been successfully applied to spatial resolution magnification of RGB images [14, 15, 16]. A straightforward idea to perform HSI SR is directly applying such networks to magnify either the spatial dimension of LR-HS image or spectral dimension of HR-RGB image, which we call Spatial-CNN and Spectral-CNN. Such naive approaches ignore the complementary advantage of fusing LR-HS and HR-RGB images.

In this paper, we propose spatial and spectral fusion architectures of CNN (SSF-CNN) to the fusion of LR-HS and HR-RGB images for HSI SR. Our SSF-CNN architecture take the concatenated cubic data of HR-RGB image and upsampled LR-HS images as input to simultaneously learn spectral attribute in LR-HS image and spatial context in HR-MS image for achieving more robust HR-HS image estimation. We further add shorter connections between the input layer and learned feature map layers to concatenate the partial data (HR-RGB image) to feature maps, which we call Partial Dense Connected SSF-CNN (PDCon-SSF). With the partial concatenated connection, the PDCon-SSF reuse the

available HR-RGB image as feature map for transferring the available maximum spatial information to the recovered HR-HS image. The main contributions of this work are two-fold: 1) we propose a novel spatial and spectral fusion architecture (SSF-CNN), which jointly exploits the narrow bands of spectral attribute in LR-HS image and the rich spatial context in HR-RGB image for HSI SR; 2) we add partially shorter connection between the input and feature map layers in SSF-CNN architecture and propose the PDCon-SSF structure, which shows even higher performance for HSI SR. Experimental results on the benchmark datasets: CAVE [17] and Harvard [18] validate that the proposed method outperforms the state-of-the-art methods in both quantitative values and visual qualities.

2. PROPOSED CNN ARCHITECTURE FOR HSI SR

The goal of HSI SR is to estimate a high resolution hyperspectral image $\mathbf{Z} \in \mathbb{R}^{W \times H \times L}$, where W and H denote the spatial dimensions and L is the spectral band number, from a LR-HS image $\mathbf{X} \in \mathbb{R}^{w \times h \times L}$ ($w \ll W$, $h \ll H$) and a HR-MS (RGB) image $\mathbf{Y} \in R^{W \times H \times l}$ ($l \ll L$). In our experiments, the available HR-MS image is a RGB image with spectral band number $l = 3$. The image formation model for depicting the relationship between the desired HR-HS and the input LR-HS images can be formulated as

$$\mathbf{X} = \mathbf{Z} *^{Spat} \mathbf{D} \downarrow + \mathbf{n} \tag{1}$$

where $\mathbf{D}$ represents a 2-dimensional (spatial) filter, $*^{Spat}$ denotes the convolutional operation in spatial domain, $\downarrow$ is the down-sampling operation, and $\mathbf{n}$ denotes the noise that follows the Gaussian distribution with zero mean value. Similarly, the image formation model for depicting the relationship between the desired HR-HS and the input HR-MS images can be formulated as

$$\mathbf{Y} = \mathbf{Z} *^{Spec} \mathbf{R} \downarrow + \mathbf{n} \tag{2}$$

where $\mathbf{R}$ represents the spectral transformation matrix (a one-dimensional spectral-directional filter) decided by camera design, which maps the HR-HS image $\mathbf{Z}$ to the HR-RGB image $\mathbf{Y}$, $*^{Spec}$ denotes the convolutional operation in spectral domain and $\downarrow$ is the down-sampling operation.

To apply CNN on the problems above, the most straightforward way is to learn HR-HS images directly from the available LR-HS images, which we call *Spatial-CNN*. For RGB images, such an approach is able to expand the spatial dimension by no more than 8. For HSI SR problem, a much larger expanding rate is always desired (e.g., 32) due to very low resolution of original LR-HS images, so Spatial-CNN may show degenerated results for HSI SR problem. With only HR-RGB image as input, it is also possible to design a CNN architecture to expand the spectral resolution to produce an HR-HS image, which we call *Spectral-CNN*. Spectral-CNN neglects the hyperspectral attribute (the relation between narrow band spectra) in HS images, so it may also show unsatisfactory results. These issues motivate use to develop CNN architectures for simultaneously taking consideration of the available LR-HS and HR-MS images, and combine spectral attribute in LR-HS image and spatial context in HR-MS image to estimate a more robust HR-HS image.

2.1. The baseline CNN architecture

Due to the promising results and design simplicity, we develop the Baseline Upsampling Network (BUN) based on CNN structure for RGB image superresolution [14, 19]. BUN mainly consists of three convolutional layers and they perform three operations to map from LR images to HR images following the schematic concept in sparse coding-based SR: patch extraction and representation, non-linear mapping, and reconstruction [14, 19]. Patch extraction and representation extracts some overlapping patches from the input image, and explains each patch as a high dimensional vector. The convolution layers in CNN acts as a non-linear function which maps a high-dimensional vector (the representation of the patches in the input) to another high-dimensional vector (the feature map in the middle-layer of CNN). Reconstruction process combines the mapped CNN features into the final HR image.

BUN share similar structures as SRCNN [14, 19], as shown in the top-row of Fig. 1. The original SRCNN uses the Y-component only in the 'Ycbcr' color space as input to predict the HR Y-image, and combines the bicubic upsampled 'cb' and 'cr' components to reconstruct the HR RGB image. The spatial filters in three convolutional layers of the SRCNN use sizes of $9 * 9$, $1 * 1$, $5 * 5$, respectively. But we make several modifications to make it suitable for the HSI SR problem. Since HSI SR attempts to recover high resolution in not only spatial but also spectral-domain (maintain the reliability of the spectral response), we use all spectral bands of the available LR images in spatial or spectral domain (the LR-HS or HR-RGB image) as input instead of the illumination image (Y-component) only like in the SRCNN [14]. With the increasing of the channel number in the input, we set the spatial filter sizes in three convolutional layers as $3*3$, $3*3$, $5*5$ with full connection in channel for the BUN of our HSI SR CNN architectures.

2.2. The variant CNN architecture of HSI SR

With BUN available, the most straightforward way to apply BUN for HSI SR is to learn the HR-HS image $\mathbf{Z}$ directly from the available LR-HS image $\mathbf{X}$, namely Spatial-CNN, or to learn $\mathbf{Z}$ from the HR-RGB image $\mathbf{Y}$, namely Spectral-CNN. However, Spatial-CNN and Spectral-CNN, perform upsampling in only one dimension of the data, while we have both $\mathbf{X}$ and $\mathbf{Y}$ as input. Next, we introduce how to jointly take advantages of $\mathbf{X}$ and $\mathbf{Y}$ using BUN.

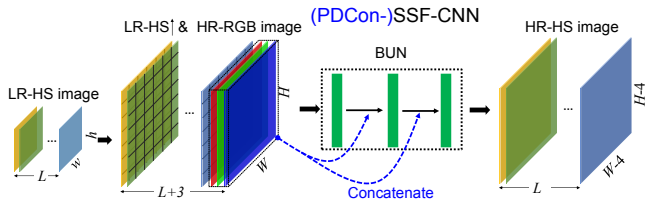

Fig. 1. The network architectures of the proposed SSF and PDCon-SSF CNN. Via adding the shortcut connection (the dot line in 'blue color') between the available HR-RGB image and the learned feature, we construct the PDCon-SSF architecture from SSF CNN.

SSF-CNN: SSF-CNN first combines the bicubic upsampled LR-HS image with the same spatial resolution as the HR-RGB image (from spatial size $w \times h$ to $W \times H$) as

$$\mathbf{Z}_{\mathbf{b}}^{\mathbf{X}} = \mathbf{X}_b \uparrow \tag{3}$$

and it then concatenates all bands (L bands) of the upsampled LR-HS image with the available HR-RGB image to form a cubic data with $L + 3$ channels as the input. The spatial sizes of the convolutional layer filters in our CNN architecture are the same as described in the above section, and the spectral sizes are the channel lengths of its inputs. The architecture of the SSF-CNN is shown Fig. 1 excepting the 'blue' color part. Since the SSF-CNN use the concatenated upsampled LR-HS and HR-RGB images as input, the first operation in this CNN architecture is patch/spectral extraction and representation instead of the spatial structure representation for the Y-component input in the original SRCNN. Because of the increased channel number, the patch/spectral representation in SSF-CNN would lead to high dimension using the spatial filter size: 9×9 of the first convolutional layer in the original SRCNN, and thus we reduce it to 3×3 while increase the spatial filter size from 1×1 to 3×3 in the second convolutional layer for taking considering of more spatial context.

PDCon-SSF: Recent CNN work incorporates shorter connections between layers for more accurate and efficient training of substantially deeper architectures such as ResNets [20] and Highway Networks [21], or exploits concatenation between different layer for information and feature reuse such as DenseNet [22], which manifest considerable improvements in different applications.

For HSI SR problem, since the available HR-RGB image has the same high spatial resolution as the output, while the expanding factor in spectral-domain is much smaller (about 10 from 3 to 31) than those in spatial-domain (32 times from 16/32 to 512/1024 in horizontal and vertical directions, respectively), we concatenate the available HR-RGB image (a part data of the input: Partial) to the outputs of the Conv and RELU blocks (Densely) in the CNN structure for transferring the available maximum spatial information, and name this new CNN architecture as PDCon-SSF. In PDCon-SSF-CNN, we add shortcut connections between the input HR-RGB image and the output feature maps of the first and second Conv/RELU layers for maximum spatial information transferring, and thus the spatial size of the HR-RGB image and the map feature in the first and second Conv/RELU layers should be same for concatenating in the channel direction. For maintain the same spatial size of the input and outputted feature maps in the first and second Conv/RELU layers, we set padding parameter as 1 with kernel size parameter 3 of the convolutional filters, and then the concatenation of the HR-RGB image to the feature maps leads to +3 channels of cubic data for the processing of next layer. The flowchart of the proposed PDCon-SSF-CNN architecture is shown in Fig. 1.

Table 1. The average and standard deviation of RMSE, PSNR, SAM and ERGAS using different CNN models on CAVE database.

	Spa.-CNN	Spec.-CNN	SSF-CNN	PDCon-SSF
RMSE	24.49±9.35	5.73±2.20	2.20±0.73	**2.18±0.75**
PSNR	20.97±3.53	33.63±3.70	41.85±3.48	**41.93±3.57**
SAM	17.61±6.30	8.40±2.37	4.39±1.42	**4.38±1.39**
ERGAS	2.54±0.89	0.57±0.37	0.23±0.11	**0.22±0.10**

3. EXPERIMENTAL RESULTS

We evaluate the proposed approach using two publicly available HS databases: CAVE dataset [17] with 32 indoor images, and Harvard dataset [18] with 50 indoor and outdoor images recorded under daylight illumination. The dimensions of the images from CAVE dataset are 512×512 pixels, with 31 spectral bands of $10nm$ wide, covering the visible spectrum from 400 to $700nm$, while the images from Harvard dataset have the dimensions of 1392×1040 pixels with 31 spectral bands of width $10nm$, ranging from 420 to $720nm$, we extract the top left 1024×1024 pixels in our experiments. We treat the original images in the databases as ground truth $\mathbf{Z}$, and downsample them by a factor of 32 to create 16×16 images using bicubic interpolation. The observed HR-RGB images $\mathbf{Y}$ are simulated by integrating the ground truth over the spectral channels using the spectral response $\mathbf{R}$ of a Nikon D7006 camera. We have randomly selected 20 HSIs from CAVE database to train CNN models, and the remaining 12 images are used for validation of the performance of the proposed CNN method. For Harvard database, 10 HSIs have been randomly selected for training, and the remaining 40 HSIs are used as test for validation. To evaluate the quantitative accuracy of the estimated HS images, four objective evaluation metrics including root-mean-square error

Table 2. The average and standard deviation of RMSE, PSNR, SAM and ERGAS using our proposed method and the state-of-the art methods of CSU [12] and NSSR [11] on both CAVE and Harvard datasets.

	CAVE dataset			Harvard dataset			
Methods	CSU [12]	NSSR [11]	PDCon-SSF	CSU [12]	NSSR [11]	SSF	PDCon-SSF
RMSE	2.97±1.09	2.37±0.91	**2.18±0.76**	1.93±1.04	**1.72±0.93**	1.76±1.00	1.74±0.95
PSNR	39.28±3.49	41.24±3.50	**41.93±3.57**	43.40±4.00	**44.35±3.84**	44.31±4.28	44.33±4.18
SAM	5.78±2.54	5.14±1.53	**4.38±1.39**	2.95±1.07	3.06±1.06	3.05±1.12	**3.03±1.15**
ERGAS	0.33±0.18	0.25±0.11	**0.22±0.10**	0.25±0.21	0.22±0.17	0.22±0.18	**0.20±0.18**

(RMSE), peak-signal-to-noise ratio (PSNR), relative dimensionless global error in synthesis (ERGAS) [23], and spectral angle mapper (SAM) [9] are used.

We compare the performance of four types of CNNs (Spatial-CNN, Spectral-CNN, SSF-CNN, and PDCon-SSF) for HSI SR. The average and the standard deviation of RMSE, PSNR, SAM and ERGAS of the 12 test images in CAVE database are shown in Table 1, which manifests much better results of the Spectral-CNN than Spatial-CNN and more performance improvement using SSF-CNN and PDCon-SSF CNN models. One recovered HS image example and the corresponding residual images with the ground-truth HR images from CAVE database is visualized in Fig. 2 using different CNN models. Since the significant performance improvement of SSF-based network over the Spatial-CNN and the Spectral-CNN has been verified, for Harvard dataset we only train the SSF-CNN and PDCon-SSF models. The average and the standard deviation of RMSE, PSNR, SAM and ERGAS of the 40 test images in Harvard database are shown in Table 2, which shows the learned SSF-CNN and PDCon-SSF models even with 10 Harvard training images only can give promising recovery performance.

We then compare with state-of-the-art HSI SR methods considering fusion of LR-HS and HR-MS images. The fusion of the LR-HS and HR-MS images have been widely explored and various methods [24, 25, 10, 13, 9, 12, 11] have been developed where couple spectral unmixing (CSU) [11] and Non-Negative Structured Sparse Representation (NSSR) [11] manifest impressive performance compared with other fusion approaches, so we only show comparison results with CSU [12] and NSSR [11] via rerunning the released source codes in [12, 11].The compared results are shown in Table 2.

In this experiment, PDCon-SSF shows the best performance on the 12 test samples of CAVE dataset, and provides the similar performance with NSSR [11] on the 40 test samples of Harvard dataset. It should be noted that we only used 10 training images (small-scale training images) randomly selected from Harvard dataset for learning CNN model parameters, and the PDCon-SSF-CNN manifests comparable average performance with NSSR [11], which have been proven to achieve the significant performance improvement [11] compared with all other state-of-the-art HSI SR methods. Via increasing the images number to include more varieties of scenes for training, we believe that more generalized CNN model can be learned and the performance of HSI SR is prospected to be further improved.

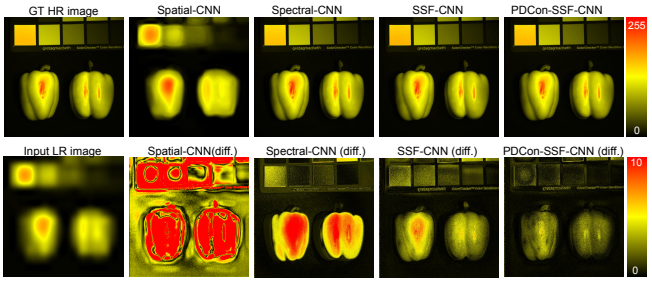

Fig. 2. An example of the reconstructed HR-HS image with the expanding factor as 32 in the spatial domain using 4 types of CNNs: Spatial-CNN, Spectral-CNN and our proposed spatial and spectral fusion CNNs: SSF-CNN and PDCon-SSF-CNN.

4. CONCLUSIONS

We proposed spatial and spectral fusion CNN (SSF-CNN) for HS image super-resolution. Instead of taking only spatial or spectral data as input, the proposed SSF-CNN concatenated the upsampled LR-HS image and the HR-RGB image to learn the high-resolution image in both spatial and spectral domains, which simultaneously took advantages of spectral attribute in LR-HS image and spatial context in HR-RGB image to estimate a more robust HR-HS image. In order to efficiently transfer the HR spatial information from the HR-RGB image to the recovered HR-HS output, we added shorter connections between the input layer and other feature map layers to concatenate the partial data (HR-RGB image) with feature maps as Partial Dense Connected SSF-CNN (PDCon-SSF). Comprehensive experiments on two public HS datasets validated that SSF-CNN and PDCon-SSF achieved better performances than state-of-the-art methods.

5. REFERENCES

[1] M. Fauvel, Y. Tarabalka, J. Benediktsson, J. Chanusssot, and J. Tilton, "Advances in spectral-spatial classification of hyperspectral images," *Proc. IEEE*, vol. 101, no. 3, pp. 652–675, 2013.

[2] M. Uzair, A. Mahmood, and A. Mian, "Hyperspectral face recognition using 3d-dct and partial least squares," *BMVC*, pp. 57.1–57.10, 2013.

[3] D. Zhang, W. Zuo, and F. Yue, "A comparative study of palmprint recognition algorithm," *ACM Comput. Aurv.*, vol. 44, no. 1, pp. 2:1–2:37, 2012.

[4] H.V. Nguyen, A. Benerjee, and R. Chellappa, "Tracking via object reflectance using a hyperspectral video camera," *CVPRW*, pp. 44–51, 2010.

[5] Y. Tarabalka, J. Chanusssot, and J. Benediktsson, "Segmentation and classification of hyperspectral images using minimum spanning forest grown from automatically selected markers," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 40, no. 5, pp. 1267–1279, 2010.

[6] Y. Zhou, H. Chang, K. Barner, P. Spellman, and B. Parvin, "Classification of histology sections via multispectral convolutional sparse coding," *CVPR*, pp. 3081–3088, 2014.

[7] J.M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. M. 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.

[8] N. Akhtar, F. Shafait, and Mian A, "Sungp: A greedy sparse approximation algorithm for hyperspectral unmixing," *ICPR*, pp. 3726–3731, 2014.

[9] Q. Wei, J. Bioucas-Dias, N. Dobigeon, and J.Y. Toureret, "Hyperspectral and multispectral image fusion based on a sparse representation," *IEEE Trans Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3658–3668, 2015.

[10] N. Akhtar, F. Shafait, and A. Mian, "Sparse spatio-spectral representation for hyperspectral image super-resolution," *ECCV*, pp. 63–78, 2014.

[11] W.S. Dong, F.Z. Fu, G.M. Shi, X. Cao, J.J. Wu, G.Y. Li, and X. Li, "Hyperspectral image super-resolution via non-negative strutured sparse representation," *IEEE Transaction on Image Processing*, vol. 25, no. 3, pp. 2337–2352, 2016.

[12] C. Lanaras, E. Baltsavias, and K. Schindler, "Hyperspectral super-resolution by coupled spectral unmixing," *ICCV*, pp. 3586–3595, 2015.

[13] E. Wcoff, T.H. Chan, K. Jia, W.K. Ma, and Y. Ma, "A non-negative sparse promoting algorithm for high resolution hyperspectral imaging," *ICASSP*, pp. 1409–1413, 2013.

[14] C. Dong, C. C. Loy, K.M. He, and X.O. Tang, "Image super-resolution using deep convolutional networks," *EEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 38, no. 2, pp. 295–307, 2015.

[15] C. Dong, C. C. Loy, and X.O. Tang, "Accelerating the super-resolution convolutional neural network," *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[17] F. Yasuma, T. Mitsunaga, D. Iso, and S.K. Nayar, "Generalized assorted pixel camera: Post-capture control of resolution, dynamic range and spectrum," *IEEE Transaction on Image Processing*, vol. 19, no. 9, pp. 2241–2253, 2010.

[18] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," *CVPR*, pp. 193–200, 2011.

[19] Y.S. Li, J. Hua, X. Zhao, W.Y. Xie, and J.J. Li, "Hyperspectral image super-resolution using deep convolutional neural network," *Neurocomputing*, vol. 266, pp. 29–41, 2017.

[20] K.M. He, X.Y. Zhang, S.Q. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[21] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *ICML 2015 Deep Learning workshop*, 2015.

[22] G. Huang, Z. Liu, K.L. Q. Weinberger, and L. V. D. Maaten, "Densely connected convolutional networks," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[23] L. Wald, "Quality of hige resolution synthesisted images: Is there a simple criterion?," *Proc. of Fusion Earth Data*, pp. 99–103, 2000.

[24] R. Kawakami, J. Wright, Y.-W. Tai, Y. Matsushita, M. Ben-Ezra, and K. Ikeuchi, "High-resolution hyperspectral imaging via matrix factorization," *CVPR*, pp. 2329–2336, 2011.

[25] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization for hyperspectral and multispectral data fusion," *IEEE Trans Geosci. Remote Sens.*, vol. 50, no. 2, pp. 528–537, 2012.