Multi-level and Multi-scale Spatial and Spectral Fusion CNN for Hyperspectral Image Super-resolution

Xian-Hua Han
Yamaguchi University, Japan
hanxhua@yamaguchi-u.ac.jp

YinQiang Zheng
National Institute of Informatics, Japan
yqzheng@nii.ac.jp

Yen-Wei Chen
Ritsumeikan University, Japan

Abstract

Hyperspectral imaging simultaneously captures images of the same scene across many numbers of spectral channels, and has different applications from agriculture, astronomy to surveillance and mineralogy, to name a few. However, due to various hardware limitations, the current hyperspectral sensor only provides low-resolution (LR) hyperspectral images compared with the RGB images obtained from a common color camera. Thus fusing a LR hyperspectral image with the corresponding high-resolution (HR) RGB image to recover a HR hyperspectral image has attracted much attention, and is usually solved as an optimization problem with prior-knowledge constraints such as sparsity representation and spectral physical properties. Motivated by the great success of deep convolutional neural network (DCNN) in many computer vision tasks, this study aims to design a novel DCNN architecture for effectively fusing the LR hyperspectral and HR-RGB images. Taking consideration of the large resolution difference in spatial domain of the observed RGB and hyperspectral images, we propose a multi-scale DCNN via gradually reducing the feature sizes of the RGB images and increasing the feature sizes of the hyperspectral image for fusion. Furthermore, we integrate multi-level cost functions into the proposed multi-scale fusion CNN architecture for alleviating the gradient vanish problem in training procedure. Experiment results on benchmark datasets validate that the proposed multi-level and multi-scale spatial and spectral fusion CNNs outperforms the state-of-the-art methods in both quantitative values and visual qualities.

1. Introduction

Hyperspectral (HS) imaging acquires images with many narrow spectral channels of a scene via densely sampling the electromagnetic spectrum. The rich spectra greatly enrich the captured scene information and have been recently applied in many computer vision tasks, such as object recognition and classification [15, 42, 47], tracking [34], segmentation [41], medical image analysis [48], and remote sensing [7, 2], for pursuing performance enhancement. However, the high spectral resolution means that a small fraction of the overall radiant energy only can be collected for each band of narrow spectrum. To guarantee acceptable signal-to-noise ratio, photon collection has to be performed in a much larger spatial region on the sensor, and thus results in low spatial resolution in the observed HS image. The low spatial resolution generally leads to high spectral mixing of different materials in the target scene, and possibly affect the performance of scene analysis and understanding. Therefore, the reconstruction of high-resolution hyperspectral (HR-HS) image using image processing and machine leaning techniques has attracted a lot of attention.

There are mainly three research directions for HR-HS image reconstruction via: 1) spatial resolution enhancement from the observed low-resolution (LR) HS image ; 2) spectral resolution enhancement from the HR-RGB image; 3) fusion method based on the observed HR-RGB and LR-HS images of a same scene. Motivated by the success of deep convolutional neural network (DCNN) for the spatial resolution enhancement in single natural image super-resolution [12, 25], some work attempted to reconstruct HR-HS image from a single LR-HS image with DCNN architecture [31, 32], and validated feasibility for small expanding factors such as 2~4. However the spatial resolution of the observed HS image is generally much lower than the commonly available RGB image, and then large expanding factor, such as more than 10 in horizontal and vertical directions, respectively, is needed for reaching the required spatial resolution of HR-HS image in real applications. Since the RGB image can be easily collected with a low-price visual sensor, the spectral resolution enhancement for RGB-to-Spectrum reconstruction [5, 16, 6, 18, 35],

has recently became an active research line. Although the potential of HR-HS image reconstruction from a single RGB image has been validated, there has still large space for performance improvement in real applications. Fusing a LR-HS image with the corresponding HR-RGB image to obtain a HR-HS image has shown promising performance [22, 10, 21, 1, 33, 49, 8, 24]. Existing effort mainly focus on optimization based fusion methods. With the spectral decomposition model, the reconstruction errors of the spectral representation for both LR-HS and HR-MS (or HR-RGB) images [46, 27, 14, 19] are jointly minimized. Since the unknown variable number in the HR-HS image is much larger than the number measurements, different constraints such as sparsity representation [22, 14, 19, 45, 17, 4, 3, 44], spectral physical properties [27], spatial context similarity [14, 19] have been used for narrowing the solution space to provide stable reconstruction. 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. In spite of the impressive performance of DCNN in different computer vision tasks, few work investigated the fusion problem due to large structure difference in the two modalities of HR-RGB and LR-HS images [20, 11, 37]. Han etc. [20] conducted a pilot study of spatial and spectral fusion CNN with simply upsampling the LR-HS image to the spatial size of the HR-RGB image, which only consists of 3 convolutional layers based on the well known SRCNN, and manifested comparable performance compared with state-of-the-art optimization-based fusion approaches. Furthermore the simple upsampling would greatly increase the amount of date, and thus leads to high computational cost. Dian et al. [11] proposed to combine the optimization- and CNN-based methods together, which consists of three independent procedures with the optimization method as the pre- and post- processing and a plain CNN architecture as the intermediate step for recovering the residual component.

In this paper, we present a novel CNN architecture to effectively fuse the observed LR-HS and HR-RGB images for HS image super resolution. The proposed CNN architecture consists of two pathways: 1) a spatial structure reservation pathway for investigating the HR structure in the HR-RGB image; 2) a spectral reservation pathway for exploring the correlation property in spectral channels. Furthermore, the learned feature maps from two pathways can be dynamically fused in a multi-scale procedure for investigating the the correlation structure between the spectral and spatial domain. The schematic of the proposed multi-scale spatial and spectral fusion CNN (MS-SSFNet) is shown in Fig. 1. As mentioned above, the spatial expanding factor of the LR-HS image is generally large, and would lead to numerous scales and deeper architecture in the MS-SSFNet. Deeper the network, the occurrence potential of gradient vanishing problem is increased. In order to alleviating the possible gradient vanish problem in training procedure, we propose to integrate multi-level cost functions for effectively training the proposed MS-SSFNet.

The main contributions of this work are two-fold:

- We propose a novel multi-scale spatial and spectral fusion architecture (MS-SSFNet), which can efficiently explore the narrow bands of spectral attribute in LR-HS image with the spectral reservation pathway and the rich spatial context in HR-RGB image with spatial structure reservation pathway for HSI SR. The fusion architecture for the learned feature maps jointly exploits the spectral and spatial correlation structure.

- We integrate multi-level cost functions for alleviating the gradient vanish problem in training procedure of deep network architecture. We divide the proposed MS-SSFNet into several levels, and construct the intermediate cost function in each level, which are combined for formulating as the final objective function in the network training procedure.

Experimental results on the benchmark datasets: Harvard [4], NUS and ICVL validate that the proposed method outperforms the state-of-the-art methods in both quantitative values and visual qualities.

The rest of the paper is organized as follows. We firstly review the related literature of HSI SR in Section 2, and then describe the proposed MS-SSFNet and the multi-level weighted objective function for network training in Section 3. Experimental evaluations are conducted in Section 4, and finally Section 5 concludes the paper.

2. Related Work

The high-resolution cubic data in both spatial and spectral domains is difficult to achieve due to technique and budget constraints [22], which motivates research attentions for generating HR-HS images via fusing HR-RGB and LR-HS images using image processing and machine learning techniques. Particularly in remote sensing field, a high resolution single-channel black-and-white ('panchromatic') image is usually available accompanying with the low resolution multi-spectral or HS image and the fusion of these two images is generally known as the pan-sharpening technique [10, 21, 1, 33, 49]. Popular approaches focused on reliable illumination restoration based on intensity substitution and projection with the sue saturation and principle component analysis [5, 16]. Generally this improves the spatial resolution of the hyperspectral image, however unavoidably causes spectral distortion [8].

Many HSI SR methods based on matrix factorization, spectral unmixing, and sparse representation, which are

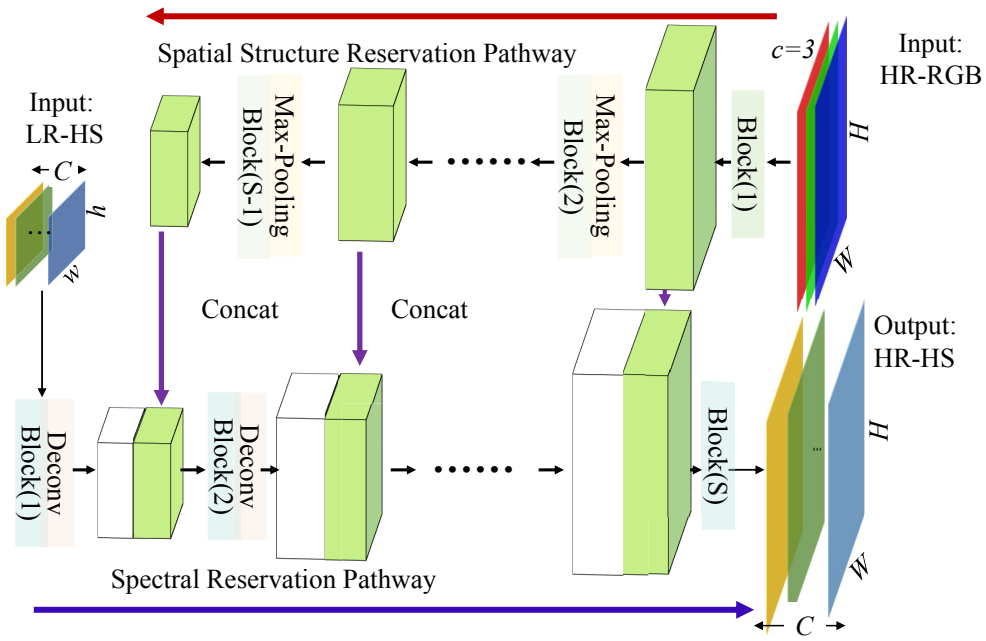


Figure 1. The proposed multi-scale spatial and spectral fusion CNN architecture.

mainly motivated by the fact that the HS observations can be represented by the basis of reflectance functions (the spectral response of the pure material) and their corresponding sparse coefficients denoting the fractions of each material on each location, has been actively investigated [24, 46, 14, 19]. Yokoya et al. [46] proposed a coupled non-negative matrix factorization (CNMF) method motivated by the prior knowledge of the non-negativity of spectral response and the spectral decomposability, which may lead to non-unique solution [29]. Lanaras et al. [27] integrated coupled spectral unmixing strategy into HSI SR, and applied the proximal alternation linearized minimization to optimize, which requires the initial points of the two decomposed reflectance functions and the endmember vectors with similar constraints. These methods require the number of pure materials in the observed scene to be smaller than the spectral band number, which does not always meet the real application.

Motivated by the success of the sparse representation in natural image analysis, the sparsity promoting approaches without explicit physical meaning constraints on the basis, which thus permits over-complete basis, have been applied for HSI SR [45, 17, 14, 19]. There are many methods using a joint sparse representation for approximating the local structure in each individual band [17], and sparse spectral representation that encodes the pixel spectrum independently instead of the local structure [4, 3]. More recently, Dong et al. [14] investigated a non-negative structured sparse representation with context similarity constraint and showed state-of-the-art performance. These methods manifested large improvement, but the performance largely relies on pre-defined constraints, which limit their wide applicability.

Deep convolutional neural networks (CNNs) have recently shown great success in various image processing and computer vision applications, such as image classification, object detection and segmentation [40, 36], face recognition [39], image denoising [30]. CNN has also been applied to RGB image super-resolution and achieved promising performance. Dong et al. [12] proposed a three-layer CNN architecture (SRCNN), which demonstrates about 0.5db-1.5db improvement and much lower computational cost compared with the popularly used sparse-based methods, and they further extended SRCNN to be capable of dealing with the available LR images without upsampling as input (Fast SRCNN) [13]. Kim et al. [25] exploited a very deep CNN architecture based on VGG-network [38], and focused on learning the missing high-frequency image (residual image) for speeding up the training procedure. Ledig et al. [28] combined GAN for estimating much sharper HR image. For applying CNN to HSI SR, Li et al. [31] applied similar structures of SRCNN to super-resolve HSI only from the LR-HS image. The CNN architectures take only the LR image as input, and the expanding factor of resolution enhancement is theoretically limited to be lower than 8 in both height and width. Recently,it attracts hot attention exploring CNN-based method with variant backbone architectures to expand the spectral resolution with only HR-RGB image as input [5, 18, 35], which is called RGB-to-hyperspectral reconstruction. Although the CNN-

based RGB-to-hyperspectral reconstruction manifested the potential of HR-HS image recovery, it easily results in spectral distortion and thus there are still large space for performance improvement. In order to exploit both information in the available HR-RGB image and LR-HS image, Han et al. [20] conducted a pilot study of spatial and spectral fusion CNN (SSF-CNN) with a baseline CNN architecture of 3-convolutional layers, which simply upsamples the LR-HS image to the same spatial size of the HR-RGB image for concatenation. Even though SSF-CNN provided comparable performance of the reconstructed HR-HSI with state-of-the-art optimization-based fusion methods, the simple upsampling of the LR-HS image would greatly increase the input data amount, which will lead to heavy computational burden. Dian et al. [11] exploited a deep hyperspectral image sharpening method (DHSIS), which is a combined strategy of the optimization- and CNN- based methods for HR-HS reconstruction. Therein, the HR-HS image is firstly optimized by minimizing the reconstruction errors with the observed HR-RGB and LR-HS images using an up-sampled LR-HS image as the initial state, which can be called as optimization-based pre-processing, and then the obtained HR-HS image in the first step is inputed to a plain network with 16 convolutional layers for estimating the residual components (the residual image between the network input and the ground-truth image). Finally, an optimization method is explored again for refining the reconstructed HR-HS image from the CNN network. This study proposes a novel CNN architecture to effectively fuse the observed LR-HS and HR-RGB images for HS image super resolution in a more robust manner.

3. Proposed Method

3.1. Problem Formulation

Let $\mathbf{Y} \in R^{W\times H\times 3}$ and $\mathbf{X} \in \mathbb{R}^{w\times h\times C}$ ($w \ll W$, $h \ll H$) denote the input HR-RGB image and LR-HS image, respectively, where W (w), H (h) are the width and height of the input image $\mathbf{Y}$ ($\mathbf{X}$), C is the spectral channel number of the LR-HS image. The goal of HSI SR is to estimate a HR-HS image $\mathbf{Z} \in \mathbb{R}^{W\times H\times C}$ from the observed LR-HS image $\mathbf{X}$ and the HR-RGB image $\mathbf{Y}$. For simplification, we consider the spatial upscale factor as 2^S, that is $W = 2^S w$, $H = 2^S h$. The image formation model for depicting the relationship between the desired HR-HS and the input LR-HS images can be formulated as

$$\mathbf{X} = \mathbf{Z} *^{Spat} \mathbf{D} \downarrow^{2^S} + \mathbf{n} \tag{1}$$

where $\mathbf{D}$ represents a 2-dimensional (spatial) filter, $*^{Spat}$ denotes the convolutional operation in spatial domain, $\downarrow^{2^S}$ is the down-sampling operation with 2^S factor for horizontal and vertical directions, respectively. $\mathbf{n}$ 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-RGB image can be formulated as

$$\mathbf{Y} = \mathbf{Z}\mathbf{R} + \mathbf{n} \tag{2}$$

where $\mathbf{R} \in R^{C\times 3}$ represents the RGB camera spectral sensitivity decided by camera design, which maps the HR-HS image $\mathbf{Z}$ to the HR-RGB image $\mathbf{Y}$. This study explores a multi-scale spatial and spectral fusion CNN architecture to integrate the observed LR-HS image $\mathbf{X}$ and the HR-RGB image $\mathbf{Y}$, which have large structure difference, in a more robust manner.

3.2. Multi-Scale Spatial and Spectral Fusion CNN: MS-SSFNet

As described in 3.1 that the spatial structure in the observed LR-HS and HR-RGB images differs largely, and thus it is difficult to fuse the two available modalities of data to generate a robust HR-HS image. We design a multi-scale SSFnet consisting of two pathways as shown in Fig. 1: spectral reservation pathway which progressively learns the upsampled spectral-correlation feature maps in multiple scales from the observed LR-HS image and simultaneously maintains spectral correlation property, and spatial structure reservation pathway which progressively learns the down-sampled spatial-correlation features from the observed HR-RGB image. And then the upsampled spectral-correlation feature maps and the down-sampled spatial-correlation features are dynamically fused also in a multi-scale manner.

With the spatial upscaling factor 2^S, there are S blocks in the spectral and spatial structure reservation pathways, respectively. A de-convolutional layer is used between the blocks of the spatial structure reservation pathway for upsampling the outputted feature with factor 2. Let $\mathbf{X}_{s-1} \in \mathbb{R}^{2^{s-1}w\times 2^{s-1}h\times C^1_{s-1}}$ and $\hat{\mathbf{X}}_s \in \mathbb{R}^{2^{s-1}w\times 2^{s-1}h\times C^1_s}$ denote the input and output feature maps of the $s-th$ block, and the relation between $\hat{\mathbf{X}}_s$ and $\mathbf{X}_{s-1}$ is formulated as:

$$\hat{\mathbf{X}}_s = F_1^{Spec}(\mathbf{X}_{s-1}, \theta_s^1) \tag{3}$$

where the input of the first block is the observed LR-HS image, etc. $\mathbf{X}_0 = \mathbf{X}$.

Via the deconvolution layer between the s and $s+1$ blocks, the spatial size of feature map $\hat{\mathbf{X}}_s$ is enlarged from $2^{s-1}w \times 2^{s-1}h$ to $2^s w \times 2^s h$. The output $\mathbf{X}_s^{Up}$ of the $s-th$ deconvolution layer is expressed as:

$$\mathbf{X}_s^{Up} = Deconv_s(\hat{\mathbf{X}}_s, \theta_s^{1,d}) \tag{4}$$

Therefore, S blocks have $S-1$-scale up-sampled feature maps: $[\mathbf{X}_1^{Up}, \mathbf{X}_2^{Up}, \cdots, \mathbf{X}_{S-1}^{Up}]$, which are combined with the learned inherent features from the spatial reservation pathway as the input of the next block.

Similarity, the spatial structure reservation pathway includes $L-1$ blocks via removing the corresponding part with the smallest spatial size block in the spectral reservation pathway since no HR spatial structure is capable of being extracted from the HR-RGB image compared to the LR-HS image. A max-pooling layer is used between the blocks of the spatial structure reservation pathway for downsampling the outputted feature with factor 2. Let $\mathbf{Y}_{s-1} \in \mathbb{R}^{2^{S-s+1}w \times 2^{S-s+1}h \times C_{s-1}^2}$ and $\hat{\mathbf{Y}}_{s-1} \in \mathbb{R}^{2^{S-s+1}w \times 2^{S-s+1}h \times C_s^2}$ denote the input and output feature maps of the $s-th$ block, and the relation between $\hat{\mathbf{Y}}_s$ and $\mathbf{Y}_{s-1}$ is formulated as:

$$\hat{\mathbf{Y}}_{s-1} = F_s^{Spa}(\mathbf{Y}_{s-1}, \theta_s^2) \tag{5}$$

where the input of the first block is the observed HR-RGB image, etc. $\mathbf{Y}_0 = \mathbf{Y}$.

The max-pooling (MP) layer between the s and $s+1$ blocks reduces the spatial size of feature map $\hat{\mathbf{Y}}_s$ from $2^{S-s+1}w \times 2^{L-s+1}h$ to $2^{S-s}w \times 2^{S-s}h$. The output $\mathbf{Y}_s$ of the $s-th$ MP layer is expressed as:

$$\mathbf{Y}_s = MP_s(\hat{\mathbf{Y}}_{s-1}) \tag{6}$$

The multi-scale fusion of the proposed SSFNet is implemented via stacking the up-sampled feature map $\mathbf{X}_s^{Up}$ of the $s-th$ block in the spectral reservation pathway and the output feature map $\hat{\mathbf{Y}}_{S-l}$ of the $(S-s+1)-th$ block in the spatial structure reservation pathway to form the input $\mathbf{X}_s$ of the $(s+1)-th$ block, which is expressed as:

$$\mathbf{X}_s = stack(\mathbf{X}_s^{Up}, \hat{\mathbf{Y}}_{S-s}) \tag{7}$$

There are altogether $S-1$ stack fusion operations for a 2^S upscale factor in our MS-SSFNet architecture. The output $\mathbf{X}_{S-1}$ of the $(S-1)-th$ stack operation is the input of final block (the $S-th$) of the spectral reservation pathway, which reconstructs the required HR-HS image from the final fused feature map. Each block in both pathways contains two convolutional layers following a PReLU layer after each convolutional layer.

3.3. Multi-level cost functions for training MS-SSFNet

For a 2^S upscale factor, our constructed MS-SSFNet consists of $S+(S-1)$ blocks: S blocks in the spectral reservation pathway and $S-1$ blocks in the spatial structure reservation pathway. With large factor, the constructed MS-SSFNet would have deep architecture, which increases the occurrence potential of the gradient vanishing problem. As we know it is easy to simulate the low resolution image from a HR image with some simple interpolation operators such as bicubic. This study simulates several intermediate resolution HS images from the training HR-HS image.

Let divide $S-$scales of our MS-SSFNet into L levels, where each level includes S_L scales and $\sum_{l=1}^{L} S_L = S$. Any training HR-HS image $\mathbf{Z}$ can be down-sampled to generate $L-1$ intermediate HS images as:

$$\begin{aligned} \mathbf{Z}^1 =& \mathbf{Z} *^{Spat} \mathbf{D}^t \downarrow^{2^{S_L}}, \mathbf{Z}^2 = \mathbf{Z}^1 *^{Spat} \mathbf{D}^t \downarrow^{2^{S_L}}, \\ & \cdots, \mathbf{Z}^{L-1} = \mathbf{Z}^{L-2} *^{Spat} \mathbf{D}^t \downarrow^{2^{S_L}} \end{aligned} \tag{8}$$

Via adding a reconstruction layer in each S^L block group, the proposed MS-SSFNet can provide estimation $\hat{\mathbf{Z}}$ of the required HR-HS image and the estimations: $\hat{\mathbf{Z}}^1, \hat{\mathbf{Z}}^2, \cdots, \hat{\mathbf{Z}}^{L-1}$ of the intermediate HS images. The L Mean Squared Errors (MSEs) in all block groups can be calculated as:

$$\begin{aligned} MSE_0 = \sum_{n=1}^{N} \|\mathbf{Z}_n - \hat{\mathbf{Z}}_n\|_2^2, MSE_1 = \sum_{n=1}^{N} \|\mathbf{Z}_n^1 - \hat{\mathbf{Z}}_n^1\|_2^2, \\ \cdots, MSE_{L-1} = \sum_{n=1}^{N} \|\mathbf{Z}_n^{L-1} - \hat{\mathbf{Z}}_n^{L-1}\|_2^2 \end{aligned} \tag{9}$$

The combined objective function for training our constructed MS-SSFNet is formulated as:

$$\arg\min_{\theta^1, \theta^{1d}\theta^2} \sum_{l=0}^{L-1} MSE_l \tag{10}$$

The schematic concept of the multi-level cost function formulation in MS-SSFNet training procedure is shown in Fig. 2. In test procedure, the reconstruction layers for the intermediate HS images will be removed, and only the HR-HS image is recovered.

4. Experimental Results

In the following, we will first introduce the datasets used for HR-HS image reconstruction, and the metrics for quantitative evaluation. Then, we compare our method with several state-of-the-art HS image reconstruction methods.

4.1. Datasets and Metrics

We evaluate the proposed multi-level and multi-scale SSFNet on three publicly hyperspectral imaging datasets including the Harvard dataset [9], the ICVL dataset [6], and the NUS dataset [35]. The Harvard dataset consists of 50 outdoor images captured under daylight illumination. We randomly select 40 images in this dataset for training and use the rest for testing. The ICVL dataset consists of 201 images, which is by far the most comprehensive natural hyperspectral dataset. This time, we conducted experiments with the used 101 images in [6], and randomly select 71 images for training and the rest for testing. The NUS dataset

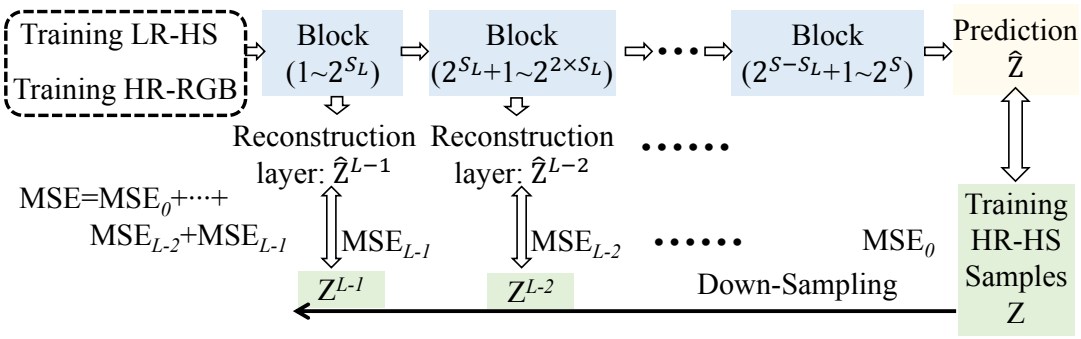


Figure 2. The schematic concept of the multi-level cost function formulation in MS-SSFNet training procedure.

Table 1. The average and standard deviation of RMSE, PSNR, SAM and SSIM using our proposed method, a comparable ResNet method, the combined optimization- and CNN- based method: DHSIS [11] and the state-of-the-art optimization-based fusion approaches: CSU [27] and NNSR [14] on all three datasets.

(a) Harvard Dataset

Methods	CSU [27]	NNSR [14]	SSF-CNN [20]	DHSIS [11]	Spectral-ResNet	SSF-ResNet	Our
RMSE	2.15±1.05	1.84±0.70	1.94±1.23	1.83±0.83	3.00±2.26	1.83±1.01	**1.54±0.61**
PSNR	42.44±4.28	43.48±3.53	43.56±4.88	43.76±4.16	40.40±5.77	44.05±4.85	**45.01±3.54**
SAM	2.79±0.66	2.83±0.70	3.14±0.97	2.64±0.71	3.83±1.54	2.47±0.65	**2.40±0.59**
SSIM	0.981±0.013	0.984±0.007	0.984±0.01	0.984±0.008	0.980±0.01	0.984±0.01	**0.986±0.006**

(b) ICVL Dataset

Methods	CSU [27]	NNSR [14]	DHSIS [11]	SSF-ResNet	Our
RMSE	0.95±0.17	0.91±0.18	0.71±0.29	0.81±0.24	**0.65±0.15**
PSNR	48.71±1.59	49.11±1.78	51.20±3.61	50.72±3.12	**51.99±2.03**
SAM	0.70±0.13	0.99±0.67	0.71±0.11	0.61±0.14	**0.53±0.11**
SSIM	0.9972±0.0012	0.9961±0.0013	0.9971±0.0013	0.9965±0.0011	**0.9979±0.0010**

(c) NUS Dataset

Methods	CSU [27]	NNSR [14]	DHSIS [11]	SSF-ResNet	Our
RMSE	1.65±0.83	1.21±0.62	1.27±0.59	1.19±0.55	**1.08±0.62**
PSNR	44.80±4.43	47.56±4.43	46.95±3.83	47.56±4.35	**48.79±4.89**
SAM	3.23±1.30	2.78±1.41	2.87±1.70	2.83±1.42	**2.71±1.55**
SSIM	0.9864±0.0136	0.9872±0.0138	0.9868±0.0068	0.9881±0.0133	**0.9883±0.0148**

contains 41 HSIs in the training set and 25 HSIs in the testing set. The HS images in all datasets have 31 spectral bands of 10 nm wide, covering the visible spectrum from 400 to 700 nm or 420 to 720 nm. We treat the original images in the datasets as ground truth $\mathbf{Z}$, and simulate to produce the observed HR-RGB images $\mathbf{Y}$ by integrating the ground truth over the spectral channels using the spectral response $\mathbf{R}$ of a Nikon D700 camera and the LR-HS images $\mathbf{X}$ by down-sampling operation with scale factor: 16 for both horizontal and vertical directions.

Four image quality metrics are utilized to evaluate the performance of our proposed method, including root-mean-square error (RMSE), peak-signal-poise-ratio (PSNR), structural similarity (SSIM) [43], and spectral angle mapping (SAM) [26]. RMSE, PSNR and SSIM are calculated on each 2D spatial image, which measure the spatial fidelity between the recovered HSI and the ground truth. SAM is calculated on the $1-D$ spectral vector, which shows the spectral fidelity. Smaller values of RMSE and SAM suggest better performance, while a larger value of PSNR and SSIM implies better performance.

4.2. Implementation Detail

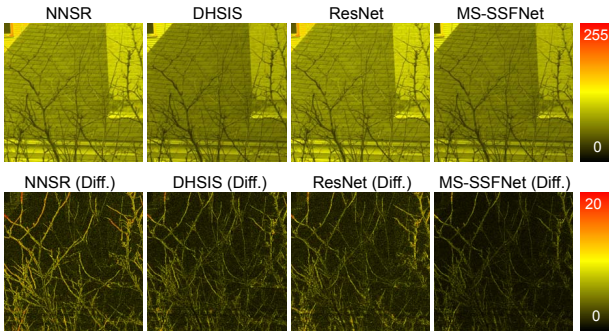
For training sample preparation, we uniformly extract the patches with the size of $8\times8\times31$ and the stride of 2 from the generated LR-HS images, the corresponding patches

with the size of $127 \times 127 \times 3$ from the generated HR-RGB images, and the corresponding ground truth with the size of $127 \times 127 \times 31$ from the HR-HS images. In addition, we divide the proposed MS-SSFNet of the upscale factor 16 into 2 levels, and generate the intermediate downsampled HS image $\hat{\mathbf{Z}}^1$ with factor 4. Therefore, the corresponding patches with the size of $31 \times 31 \times 31$ are also extracted as the intermediate ground-truth samples. Our network, implemented with Caffe [23], is trained from scratch, using the Adam optimizer. We use a minibatch size of 16 in training procedure, and train the network for 2000 epochs for all three datasets. The model parameters are initialized according to Gaussian distribution with standard deviation 0.001.

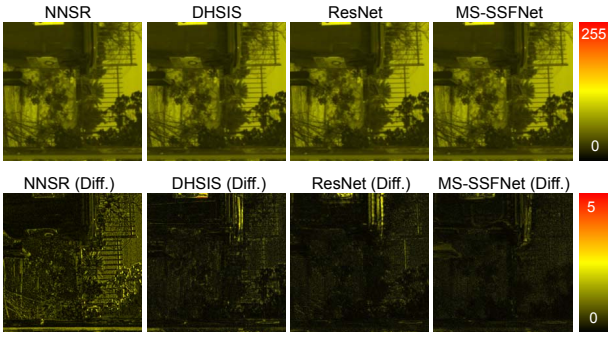
4.3. Comparison with state-of-the-art fusion methods

As we introduced above, our proposed method aims at fusing the observed LR-HS and HR-RGB images to generate a HR-HS image, and then we compare the performance of our proposed multi-level and multi-scale SSFNet with the state-of-the-art fusion methods. The state-of-the-art fusion methods for HS image super resolution mainly contains two categories: optimization-based and CNN-based strategies. There have been many recently-proposed optimization based fusion methods, including Coupled Non-negative Matrix Factorization (CNMF) method [46], Sparse Non-negative Matrix Factorization (SNNMF) method [44], Generalization of Simultaneous Orthogonal Matching Pursuit (GSOMP) method [49], Bayesian sparse representation (BSR) method [4], Couple Spectral Unmixing (CSU) method [27], and Non-Negative Structured Sparse Representation (NNSR) method [14]. Since CSU and NNSR manifest relatively larger advantage over other existing optimization-based methods [46, 44, 46, 4], we only provide the compared results of our proposed method with the CSU and NNSR optimization-based fusion methods in Table 1. From Tabel 1, we observe that for all quantitative metrics our approach can greatly improve the performances for all three datasets.

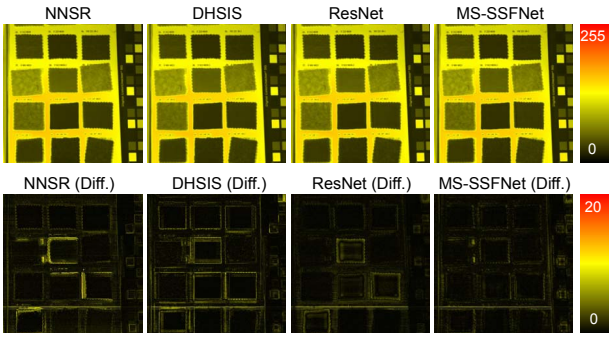
In recent year, CNN-based HR image reconstruction has been actively investigated, which mainly focused on the spectral resolution enhancement from a single RGB image, and has proven the potential of this RGB-to-hyperspectral reconstruction. Because the difference of spatial resolution in the observed LR-HS and HR-RGB images are considerable large, few CNN-based fusion work is explored. A pilot study proposed a simple spatial and spectral fusion CNN (SSF-CNN) [20], which adopted three convolutional layers based on the the SRCNN model for natural image super resolution. DHSIS method [11] also explored the hyperspectral reconstruction via integrating the optimization- and CNN- based methods. We re-conducted experiments with the SSF-CNN architecture and the DHSIS method [11] in



(a) An image from Harvard dataset



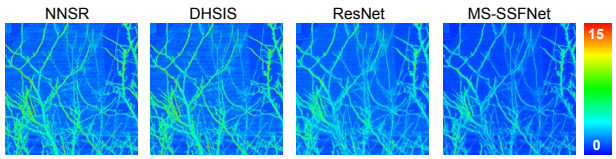
(b) An image from ICVL dataset



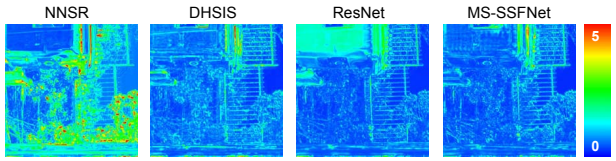
(c) An image from NUS dataset

Figure 3. An image example from each of three datasets. Te recovered images by a state-of-the-art optimization fusion method: NNSR [14], the comparable ResNet-based fusion network, the combined optimization- and CNN-based method: DHSIS [11] and our proposed MS-SSFNet are give in the first row. The error images between the recovered images and the ground-truth are provided in the second row.

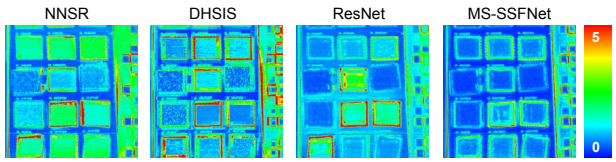
the same conditions (optimization method of network training, epoch number, training and testing samples etc.) as in our proposed MS-SSFNet. The compared results on Harvard dataset is shown in Table 1(a). Our proposed MS-SSFNet contains several blocks while the SSF-CNN is constitute of three convolutional layers only. In order to provide fair comparison, we also design a much deeper CNN architecture, which consists of five blocks of residual connection



(a) An image from Harvard dataset



(b) An image from ICVL dataset



(c) An image from NUS dataset

Figure 4. The visualized images with the calculated Sam values (Anger degree in the range [0, 180]) as the intensity from our proposed method, the comparable ResNet-based fusion method, the optimization-based fusion approach: NNSR [14], and the combined optimization- and CNN-based method: DHSIS [11]. The Sam values are computed between the pixel spectral vectors of the recovered image and the ground-truth image.

structure with two convoutional layers in each block, called as ResNet. The designed ResNet has similar depth with our proposed MS-SSFNet but with only one pathway for taking the concatenated up-sampled LR-HS image and HR-RGB image (named as SSF-ResNet) or the HR-RGB image only (named Spectral-ResNet) as input. The compared results on Harvard dataset are provided in Table 1(a). From Table 1(a), we observe that DHSIS method [11] and the SSF-ResNet manifest better performance than the Spectral-ResNet and the SSF-CNN [20]. Thus, next for ICVL and NUS datasets, we only implemented the CNN-based fusion methods of the DHSIS method, our designed SSF-ResNet and our proposed MS-SSFNet. The compared performances are given in Table 1(b) and (c) for ICVL and NUS datasets, respectively. Table 1 manifests that our proposed method provides the best performance on all quantitative metrics for three datasets.

To visualize the experimental results for different fusion methods including the optimization-based fusion method: NNSR [14], DHSIS [11], the SSF-ResNet and our proposed MS-SSFNet, a representative recovered HS image from each of three datasets are shown in Fig. 3. The recovered results by NNSR, DHSIS, ResNet-based Fusion and our proposed MS-SSFNet are shown in the first row, and their error images are provided in the second row. All results are the $25-th$ band of the hyperspectral images. The error images are the absolution errors between the ground truth and the recovered results. We can observe that the error images from our method have much smaller magnitudes than those from the NNSR method [14], DHSIS [11] and ResNet-based fusion approach for all images from different datasets, which validates that our method can provide higher spatial accuracy. In addition, we also calculated the Sam values of all pixels to measure spectral distortion for the representative images in three datasets, and normalized the Sam values to the anger degree range $[0, 180]$. The visualized images with the calculated Sam values as intensity are shown in Fig. 4. Small magnitudes in the visualized Sam images mean the small anger degrees between the recovered spectra and the ground-truth spectra. From Fig. 4, we can see that the Sam images with our proposed MS-SSFNet manifest much smaller values for most pixels in all images than the NNSR [14], DHSIS [11] and ResNet-based fusion approaches, which verifies that our method can obtains higher spectral fidelity.

5. Conclusions

In this paper, we have presented an effective CNN-based method for fusing the observed LR-HS and HR-RGB images to reconstruct a HR-HS image. Since the structure in the two observed modalities of data: LR-HS and HR-RGB images have very large difference, it is difficult to effectively combine them in one CNN stream. This study proposed a multi-scale spatial and spectral fusion CNN, which consists of two pathways: 1) a spatial structure reservation pathway for investigating the HR spatial structure in the HR-RGB image; 2) a spectral reservation pathway for exploring the correlation property in spectral channels. The proposed MS-SSFNet gradually learn the high-resolution features from the LR-HS image and spatial-reduced features from the HR-RGB image to conduct multiple fusions for exploring the spatial and spectral correlation. Furthermore, we integrated multi-level cost functions in training procedure to alleviate the gradient vanishing problem, which is possibly appeared due to the long forward and backward propagation chains in the MS-SSFNet for large spatial upscale factor. Experimental results showed that our method can provide substantial improvements over the current state-of-the-art methods in terms of both objective metric and subjective visual quality.

6. Acknowledge

This study is supported in part by the JSPS KAKENHI under Grand No. 19K20307.

References

[1] B. Aiazzi, S. Baronti, F. Lotti, and M. selva. A comparison between global and context-adaptive pansharpening of multispectral images. *IEEE Geosci. Remote Sens. Lett.*, 6(2):302–306, 2009. 2

[2] N. Akhtar, F. Shafait, and M. A. Sungp: A greedy sparse approximation algorithm for hyperspectral unmixing. *ICPR*, pages 3726–3731, 2014. 1

[3] N. Akhtar, F. Shafait, and A. Mian. Sparse spatio-spectral representation for hyperspectral image super-resolution. *ECCV*, pages 63–78, 2014. 2, 3

[4] N. Akhtar, F. Shafait, and A. Mian. Bayesian sparse representation for hyperspectral image super resolution. *CVPR*, pages 3631–3640, 2015. 2, 3, 7

[5] A. Alvarez-Gila, J. van de Weijer, and E. Garrote. Adversarial networks for spatial context-aware spectral image reconstruction from rgb. *IEEE International Conference on Computer Vision Workshop (ICCVW 2017)*, 2017. 1, 2, 3

[6] B. Arad and O. Ben-Shahar. Sparse recovery of hyperspectral signal from natural rgb images. *In European Conference on Computer Vision*, pages 19–34, 2016. 1, 5

[7] J. 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.*, 1(2):6–36, 2013. 1

[8] M. Cetin and N. Musaoglu. Merfing hyperspectral and panchromatic image data: Qualitative and quantitative analysis. *Int. J. Remote Sens.*, 30(7):1779–1804, 2009. 2

[9] A. Chakrabarti and T. Zickler. Statistics of real-world hyperspectral images. *CVPR*, pages 193–200, 2011. 5

[10] P. Chavez, S. Sides, and J. Anderson. Comparison of three different methods to merge multiresolution and multispectral data: Landsat tm and spot panchromatic. *Photogramm. Eng. Rem. S.*, 30(7):1779–1804, 1991. 2

[11] R. W. Dian, S. Li, A. Guo, and L. Fang. Deep hyperspectral image sharpening. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–11, 2018. 2, 4, 6, 7, 8

[12] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(2):295–307, 2015. 1, 3

[13] C. Dong, C. C. Loy, and X. Tang. Accelerating the super-resolution convolutional neural network. *Proceedings of European Conference on Computer Vision (ECCV)*, 2016. 3

[14] W. Dong, F. Fu, G. Shi, X. Cao, J. Wu, G. Li, and X. Li. Hyperspectral image super-resolution via non-negative structured sparse representation. *IEEE Transaction on Image Processing*, 25(3):2337–2352, 2016. 2, 3, 6, 7, 8

[15] M. Fauvel, Y. Tarabalka, J. Benediktsson, J. Chanusssot, and J. Tilton. Advances in spectral-spatial classification of hyperspectral images. *Proc. IEEE*, 101(3):652–675, 2013. 1

[16] S. Galliani, C. Lanaras, D. Marmanis, E. Baltsavias, and K. Schindler. Learned spectral super-resolution. *arXiv preprint arXiv:1703.09470*, 2017. 1, 2

[17] C. Grohnfeldt, X. X. Zhu, and R. Bamler. Jointly sparse fusion of hyperspectral and multispectral imagery. *IGARSS*, 2013. 2, 3

[18] X.-H. Han, B. Shi, and Y. Zheng. Residual hsrcnn: Residual hyper-spectral reconstruction cnn from an rgb image. *ICPR*, pages 2664–2669, 2018. 1, 3

[19] X.-H. Han, B. Shi, and Y. Zheng. Self-similarity constrained sparse representation for hyperspectral image super-resolution. *IEEE Transaction on Image Processing*, 27(11):5625–5637, 2018. 2, 3

[20] X.-H. Han, B. Shi, and Y. Zheng. Ssf-cnn: Spatial and spectral fusion with cnn for hyperspectral image super-resolution. *ICIP*, pages 2506–2510, 2018. 2, 4, 6, 7, 8

[21] R. Haydn, G. Dalke, J. Henkel, and J. Bare. Application of the ihs color transform to the processing of multisensor data and image enhancement. *Int. Symp on Remote Sens. of Env.*, 1982. 2

[22] B. Huang, H. song, H. Cui, J. Peng, and Z. Xu. Spatial and spectral image fusion using sparse matrix factorization. *IEEE Trans Geosci. Remote Sens.*, 52(3):1693–1704, 2014. 2

[23] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 7

[24] R. Kawakami, J. Wright, Y.-W. Tai, Y. Matsushita, M. Ben-Ezra, and K. Ikeuchi. High-resolution hyperspectral imaging via matrix factorization. *CVPR*, pages 2329–2336, 2011. 2, 3

[25] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 3

[26] F. Kruse, A. Lefkoff, J. Boardman, K. Heidebrecht, A. Shapiro, P. Barloon, and A. Goetz. The spectral image processing system (sips)interactive visualization and analysis of imaging spectrometer data. *Remote Sensing of Environment*, 44(2-3):145–163, 1993. 6

[27] C. Lanaras, E. Baltsavias, and K. Schindler. Hyperspectral super-resolution by coupled spectral unmixing. *ICCV*, pages 3586–3595, 2015. 2, 3, 6, 7

[28] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. *Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[29] D. D. Lee and S. H. Seung. Algorithms for non-negative matrix factorization. *NIPS*, pages 556–562, 2001. 3

[30] H. Li. Deep learning for image denoising. *International Journal Signal Processing, Image Processing and Pattern Recognition*, 7(3):171–180, 2014. 3

[31] Y. Li, J. Hua, X. Zhao, W. Xie, and J. Li. Hyperspectral image super-resolution using deep convolutional neural network. *Neurocomputing*, 266:29–41, 2017. 1, 3

[32] S. Mei, X. Yuan, J. Ji, and Y. Zhang. Hyperspectral image spatial super-resolution via 3d full convolutional neural network. *Neurocomputing*, 9(11), 2017. 1

[33] A. Minghelli-Roman, L. Polidori, S. Mathieu-Blanc, L. Loubersac, and F. Cauneau. Spatial resolution improvement by merging meris-etm images for coastal water monitoring. *IEEE Geosci. Remote Sens. Lett.*, 3(2):227–231, 2006. 2

[34] H. Nguyen, A. Benerjee, and R. Chellappa. Tracking via object reflectance using a hyperspectral video camera. *CVPRW*, pages 44–51, 2010. 1

[35] R. M. Nguyen, D. K. Prasad, and M. S. Brown. Training-based spectral reconstruction from a single rgb image. *In European Conference on Computer Vision*, pages 186–201, 2014. 1, 3, 5

[36] W. Ouyang, X. Wang, X. Zeng, and S. Qiu. Deepid-net: deformable deep convolu- tional neural networks for object detection. *Proceedings of the Computer Vision and Pattern Recognition*, pages 2403–2412, 2015. 3

[37] Y. Qu, H. Qi, and C. Kwan. Unsupervised sparse dirichlet-net for hyperspectral image super-resolution. *CVPR*, pages 2862–2869, 2018. 2

[38] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 3

[39] Y. Sun, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. *Adv. Neural Inf. Process. Syst.*, 27:1988–1996, 2014. 3

[40] C. Szegedy, S. Reed, D. Erhan, D. Anguelov, and S. Ioffe. Scalable, high-quality object detection. *arXiv: 1412.1441v3*, pages 1–10, 2015. 3

[41] 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.*, 40(5):1267–1279, 2010. 1

[42] M. Uzair, A. Mahmood, and A. Mian. Hyperspectral face recognition using 3d-dct and partial least squares. *BMVC*, pages 57.1–57.10, 2013. 1

[43] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4):600–612, 2004. 6

[44] E. Wcoff, T. Chan, K. Jia, W. Ma, and Y. Ma. A non-negative sparse promoting algorithm for high resolution hyperspectral imaging. *ICASSP*, pages 1409–1413, 2013. 2, 7

[45] Q. Wei, J. Bioucas-Dias, N. Dobigeon, and J. Toureret. Hyperspectral and multispectral image fusion based on a sparse representation. *IEEE Trans Geosci. Remote Sens.*, 53(7):3658–3668, 2015. 2, 3

[46] N. Yokoya, T. Yairi, and A. Iwasaki. Coupled nonnegative matrix factorization for hyperspectral and multispectral data fusion. *IEEE Trans Geosci. Remote Sens.*, 50(2):528–537, 2012. 2, 3, 7

[47] D. Zhang, W. Zuo, and F. Yue. A comparative study of palmprint recognition algorithm. *ACM Comput. Aurv.*, 44(1):2:1–2:37, 2012. 1

[48] Y. Zhou, H. Chang, K. Barner, P. Spellman, and B. Parvin. Classification of histology sections via multispectral convolutional sparse coding. *CVPR*, pages 3081–3088, 2014. 1

[49] R. Zurita-Milla, J. Clevers, and M. E. Schaepman. Unmixing-based landsat tm ane meris fr data fusion. *IEEE Geosci. Remote Sens. Lett.*, 5(3):453–457, 2008. 2, 7