EasyChair Preprint

№ 7028

Spatial-Spectral Joint Correction Network for Hyperspectral and Multispectral Image Fusion

Tingting Wang, Yang Xu, Zebin Wu and Zhihui Wei

November 30, 2021

Spatial Spectral Joint Correction Network for Hyperspectral and Multispectral Image Fusion *

Tingting Wang[1], Yang Xu[1], Zebin Wu[1], and Zhihui Wei[1]

[1]School of Computer Science and Engineering,
Nanjing University of Science and Technology, Nanjing, China
wangtiting@njust.edu.cn, xuyangth90@njust.edu.cn,
wuzb@njust.edu.cn, gswei@njust.edu.cn

Abstract. Hyperspectral and multispectral image (HS-MSI) fusion aims to generate a high spatial resolution hyperspectral image (HR-HSI), using the complementarity and redundancy of the low spatial resolution hyperspectral image (LR-HSI) and the high spatial resolution multispectral image (HS-MSI). Previous works usually assume that the spatial down-sampling operator between HR-HSI and LR-HSI, and the spectral response function between HR-HSI and HR-MSI are known, which is infeasible in many cases. In this paper, we propose a coarse-to-fine HS-MSI fusion network, which does not require the prior on the mapping relationship between HR-HSI and LRI or MSI. Besides, the result is improved by iterating the proposed structure. Our model is composed of three blocks: degradation block, error map fusion block and reconstruction block. The degradation block is designed to simulate the spatial and spectral down-sampling process of hyperspectral images. Then, error maps in space and spectral domain are acquired by subtracting the degradation results from the inputs. The error map fusion block fuses those errors to obtain specific error maps corresponding to initialize HSI. In the case that the learned degradation process could represent the real mapping function, this block ensures to generate accurate errors between degraded images and the ground truth. The reconstruction block uses the fused maps to correct HSI, and finally produce high-precision hyperspectral images. Experiment results on CAVE and Harvard dataset indicate that the proposed method achieves good performance both visually and quantitatively compared with some SOTA methods.

Keywords: Hyperspectral Image · Image Fusion · Deep Learning · Degradation Model

* This work was supported in part by the National Natural Science Foundation of China (61772274, 62071233, 61671243, 61976117), the Jiangsu Provincial Natural Science Foundation of China (BK20211570, BK20180018, BK20191409), the Fundamental Research Funds for the Central Universities (30917015104, 30919011103, 30919011402, 30921011209), and in part by the China Postdoctoral Science Foundation under Grant 2017M611814, 2018T110502.

1 Introduction

With the steady development of sensor technology, the quantity and expression form of information are gradually enriched. Hyperspectral remote sensing image is mainly formed by acquiring electromagnetic waves of different wavelengths which are reflected from the ground objects after processing. Thus, the hyperspectral image generally consists of tens to hundreds of wavelengths and contains rich spectral information. Using different feature signals in hyperspectral images, many computer vision tasks such as detection [16,11] and segmentation [13] can be implemented. However, due to the limitation of existing optical remote sensing systems, it is difficult to guarantee both the spectral resolution and spatial resolution of HSI. High precision HR-HSI can provide high-quality data for subsequent more complex hyperspectral image processing tasks, and it can be produced by making full use of the MSI or HSI which can be captured by existing imaging equipment. Therefore, researchers have proposed a variety of hyperspectral image fusion methods to generate accurate HR HSI.

When composed of a single band, the multispectral image is reduced to a panchromatic image [10]. Consequently, the comprehensive evaluation of HS-MS fusion can be incorporated into the system of pan-sharpening, and the methods of HS-MS fusion and pan-sharpening are convergent. Most recent HS-MS fusion methods are based on image prior models, which formulate the fusion problem as an optimization problem constrained by HRI priors. In addition, some methods exploit the low-rank and sparse properties of HSI. These methods use matrix factorization or tensor factorization to characterize HSI and address the corresponding image fusion problem.

As recent years, deep learning (DL) in inverse problem reconstruction has gradually attracted wide attention from researchers with the continuous development of neural networks. Using back propagation of neural networks and optimization algorithms, the optimization problem can be solved effectively and achieve excellent reconstruction results. Compared with conventional fusion methods, DL-based ones need fewer assumptions on the prior knowledge from the to-be-recovered HR-HSI and the network can be trained directly on a set of training data. Although the network architecture itself needs to be handcrafted, properly designed network architectures have been shown to solve many problems and achieve high performance because of the robust feature extraction capabilities of convolutional networks [6]. Hence, based on CNN and the generation mechanism, we propose a spatial-spectral joint correction HS-MS fusion network (SSJCN). The implementation of the method revolves around the following points:

1. Improving the resolution accuracy of the fused images by concatenating the degradation models and the reconstruction models.
2. The error map between the degraded image and the input data maintains the high-frequency information of the input to ensure that the network does not lose detail information during forward propagation.

The rest of this article is organized as follows. In Section 2, we present some existing methods of hyperspectral fusion. In Section 3, we introduce the detailed implementation of the proposed model. Experimental results on two publicly available datasets and comparisons with other methods are reported in Section 4. Lastly, this paper ends with the summary of Section 5.

2 Related Work

2.1 Traditional methods

Generally, the traditional approach is based on artificial priori assumptions. There are several pan-sharpening methods often assume that the spatial details of panchromatic and multispectral images are similar [9]. While some methods, such as [3,7], use sparse matrix decomposition to learn the spectral dictionary of LR-HSI, and then use the spectral dictionary and the coefficients learned from HR-MSI to construct HR-HSI. In [3], W. Dong et al. take the spatial structure into account to make full use of the priors. Also, tensor factorization-based methods have made great strides in hyperspectral image fusion problems, which treat HR-HSI as a three-dimensional tensor [8]. Although these methods are constantly evolving and have yielded positive results, the methods based on handcrafted priors are not flexible enough to adapt to different hyperspectral image structures because HR-HSI acquired from real scenes are highly diverse in both spatial and spectral terms.

2.2 Deep learning methods

Unlike traditional methods, deep learning-based fusion methods do not require building a specific priori model. Chao Dong [2] et al. proposed a three-layer super resolution model of convolutional neural networks (SRCNN) to learn the inherently unique feature relationships between LRI and HRI. SRCNN first demonstrates that the traditional sparse coding-based approach can be reformulated as a deep convolutional neural network, but the method does not consider the self-similarity of the data. Shuang Xu [15] et al. designed a multiscale fusion network (HAM-MFN), where the HSI was upscale 4 times and fused with MS images at each scale with the net going deeper. As existing imaging equipment cannot directly obtain HR-HSI, some methods use up-sampled LR-HSI or HR-MSI to simulate the target image. Based on that, Han et al. proposed spatial and spectral fusion CNN [4]. Even though this algorithm achieved better performance than state-of-the-art methods, the up-sampled images not only increased the number of pixels but also the computational complexity. Then in [5], a multi-scale spatial and spectral fusion architecture (MS-SSFNet) is proposed in order to reducing the computational complexity and alleviating the vanishing gradients problem.

3 Proposed Method

3.1 Problem Formulation

Given two input images: HR-MSI $X \in R^{c\times W\times H}$, and LR-HSI $Y \in R^{C\times w\times h}$ ($c \ll C$, $w \ll W$, $h \ll H$) where C, W and H represents the numbers of spectral bands, image width and height respectively. The purpose of hyperspectral image fusion is to produce a potential HR-HSI $Z \in R^{W\times H\times C}$ from the observed images. Usually, we describe the relationship between Z and X, Y in the following equation.

$$X = Z \times_3 P \tag{1}$$

$$Y = Z \times_1 S_1 \times_2 S_2 \tag{2}$$

Eq.(1) indicates how to obtain HR-MSI X with the spectral response operator $P \in^{C\times c}$ ($c < C$). $S_1 \in^{W\times w}$ and $S_2 \in^{H\times h}$ in Eq.(2) is for blurring HR-HSI Z (usually using Gaussian filtering) and S denotes the spatial down-sampling operator. That is, X is a down-sampling of Z in the spectral dimension while the LR-HSI Y is generated by down-sampling the HR-HSI. The proposed model estimates Z using an end-to-end mapping function $f(\bullet)$ with the network parameters as

$$\hat{Z} = f_\theta(X,\ Y), \theta = \{w_1, \ldots, w_l; b_1, \ldots, b_l\} \tag{3}$$

where $\hat{Z}$ is the reconstructed HSI by the fusion network and w_l and b_l represent the weight and bias of the lth layer.

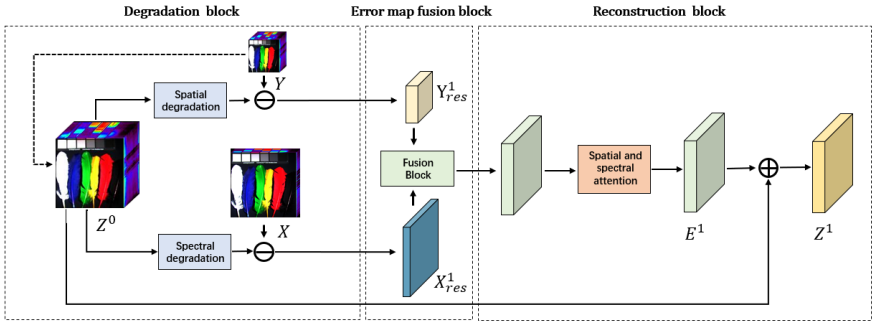

Fig. 1. The structure of the degradation block, the error map fusion block and the reconstruction block.

3.2 Degradation Block

Our network includes two inputs: LR-HSI Y and HR-MSI X. Since the HR-MSI is spectral down-sampled, it has more spatial information than LR-HSI. Correspondingly, the LR-HSI preserves more spectral information than the HR-MSI. In order to simulate the degradation model using the convolutional network, we upscale LR-HSI by bicubic interpolation to obtain input data of the same size as the HR-HSI. The result of the up-sampling is denoted by Z^0,which can be considered as a rough estimate of HR-HSI.

At first, we feed Z^0 into the network and let it pass through the spectral and spatial degradation blocks respectively, which can be expressed as

$$\hat{X}^1 = D_{spe}\left(Z^0\right) \tag{4}$$

$$\hat{Y}^1 = D_{spa}\left(Z^0\right) \tag{5}$$

Many algorithms consider the spectrum down-sampling operator P in Eq.(1) as a matrix and then HR-MSI can be calculated by simple matrix multiplication. We apply the $D_{spe}(\bullet)$ for modelling the HR-HSI spectral degradation mechanism, which is composed of a convolutional layer and an activation function layer. While the operator B and S in Eq.(2) have usually been implemented with convolution and pooling. The function $D_{spa}(\cdot)$ has the same structure as $D_{spe}(\bullet)$ and represents the non-linear mapping between LR-HSI and HR-HSI. Secondly, the spatial and spectral residuals of Z^0 are obtained by differencing the degradation results obtained in the first step with the observed data, which can be written as

$$X_{res}^1 = X - \hat{X}^1 \tag{6}$$

$$Y_{res}^1 = Y - \hat{Y}^1 \tag{7}$$

Finally, X_{res}^1 and Y_{res}^1 are used as the input of the error map feature extraction block to estimate residual map E^1 between Z^0 and Z. The implementation detail is described in section 3.3.

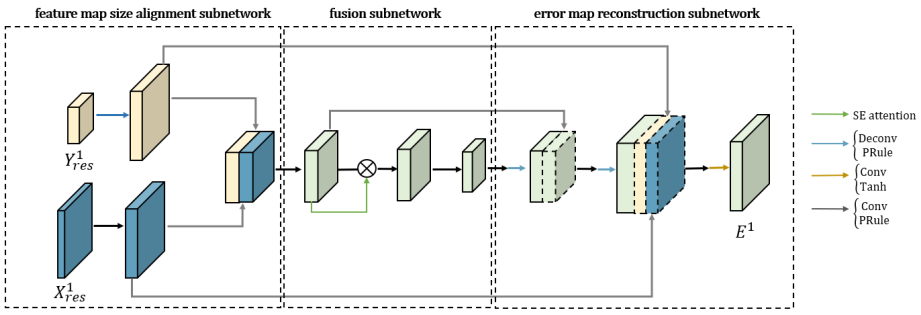

Fig. 2. The specific structure of the error map fusion block.

3.3 Error Map Fusion Block

In this section, we will specify the network structure of the error map fusion block like the Fig. 2. The X_{res}^1 and Y_{res}^1 outputted from the degradation block are used to produce the particular error map corresponding to Z^0. Degraded images $\hat{X}^1$ and $\hat{Y}^1$ retain the effective low-level semantic information of the Z^0 during the forward propagation of the network. Thus, residual data X_{res}^1 and Y_{res}^1 are more accurate for the correction of Z^0. The error map fusion process can be expressed as

$$Z^0 = \Phi(X_{res}^1 \ , \ Y_{res}^1 \ ) \tag{8}$$

Feature map size alignment subnetwork. The size of LR-HSI $\hat{Y}^1$ is smaller than the HR-MSI $\hat{X}^1$ as it is spatially down-sampled. To maintain the consistency of the feature size between the extracted features and the fusion results, we up-sample the Y_{res}^1 using deconvolution while the learnable CNN can improve the up-sampling results for each channel. And then passed it through a down-sampling layer for initial feature extraction. Meanwhile, the X_{res}^1 is passed through a low-level feature extraction block consisting of one convolution layer and one PRelu layer. After that, it was also seed to a down-sampling layer.

Fusion subnetwork. The features extracted from X_{res}^1 and Y_{res}^1 are concatenated along the spectral dimension. The fusion subnet consists of two convolution layers with separate activation functions. We incorporate a channel attention mechanism to the results of the first fusion layer to better preserve the spectral structure and reduce redundant information. The feature map is down-sampled again after this subnetwork. At this time, we have obtained the feature maps containing the spatial and spectral information simultaneously, which will be used to reconstruct the residual map of Z^0 by the subsequent up-sampling network.

Error map reconstruction subnetwork. The error map reconstruction subnetwork is composed of two sets of network structures, each comprising successive convolutional, deconvolution layers and activation functions. In an effort to take advantage of the complementary nature of the higher-level and lower-level features, jump connections are added to the features after each deconvolution. The last concatenated feature maps are then convolved in two layers to obtain the final result. In this way, the error feature map $\hat{E}^1$ is acquired from this subnetwork. $\hat{E}^1$ contains high frequency information proposed from X_{res}^1 and Y_{res}^1, and used to rebuild the final result.

3.4 Reconstruction Block

The reconstruction block refines the Z^0 with the error map $\hat{E}^1$ outputted after the first two blocks. In order to make the error maps obtained in the previous part better modify the initial hyperspectral image, we apply spatial and spectral attention model to the maps. And then $\hat{E}^1$ are added to the Z^0 as shown by the

skip connection in Fig. 1 to produce the reconstructed hyperspectral image Z^1, which can be written as

$$Z^1 = Z^0 + \hat{E}^1 \tag{9}$$

To improve the accuracy of Z^1, we refine it by one more degradation and reconstruction operation and the implementation process is the same as the three blocks above. The whole process is shown in the Fig. 3, and the final output is expressed as

$$Z^2 = Z^1 + \hat{E}^2 = Z^1 + \Phi\left(X^2_{res},\ Y^2_{res}\right) = Z^1 + \Phi\left(D_{spe}\left(Z^1\right),\ D_{spa}\left(Z^1\right)\right) \tag{10}$$

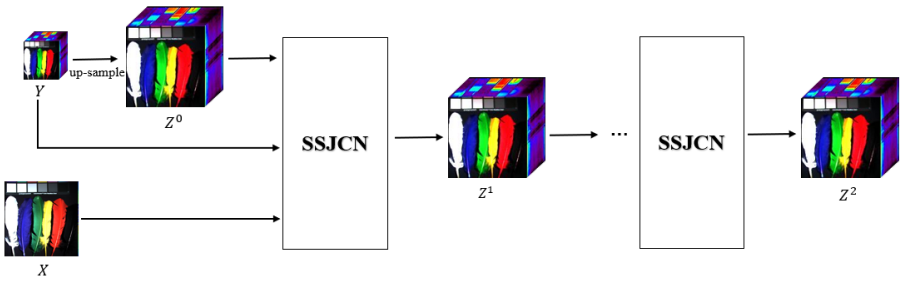

Fig. 3. The overall structure of the proposed spatial spectral joint correction network.

3.5 Loss Function

In our model, we reconstruct the HR-HSI by learning the mapping function $f_\theta(X,\ Y)$. The parameters θ are optimized by minimizing the loss between the outputs and the observed images. We choose the L1 norm function as the loss function for it is simple to implement and achieves good results in image super-resolution [17]. Thus, the loss function defined as

$$l(\theta) = \|Z - Z^2\| = \|f_\theta(X, Y) - Z^2\| \tag{11}$$

X and Y represent known LR-HSI and HR-MSI, which are obtained from the spatial and spectral down-sampling of the true value Z respectively.

4 Experiments and Analysis

4.1 Data and Experimental Setup

We conducted experiments on CAVE and Harvard dataset. The CAVE dataset consists of HR-HSI captured under 32 indoor scenes with manipulated illumination. Each HR-HSI is 512∗512∗31 in size, where 512 is the spatial size of the

image and 31 is the number of channels in the image, representing the reflectance of the material in the scene at different spectra. The Harvard database contains 50 images taken under daylight illumination, and 27 images under artificial or mixed lighting. In this experiment we use 50 images under daylight illumination. The first 20 of these HSIs are assigned as the training set, the middle 5 are used as the validation set, and the last 25 HSIs are used for testing.

We compare the proposed method with HySure [12], DHSIS [1] and DBIN [14]. HySure formulates the fusion problem as a convex optimization problem, which is solved by the split augmented Lagrange algorithm (SALSA). DHSIS optimizes the modeling results using the prior information extracted from the convolutional network, while DBIN is a network structure built entirely from convolutional layers. Fig. 4 shows the output images obtained by the several methods and corresponding error maps.

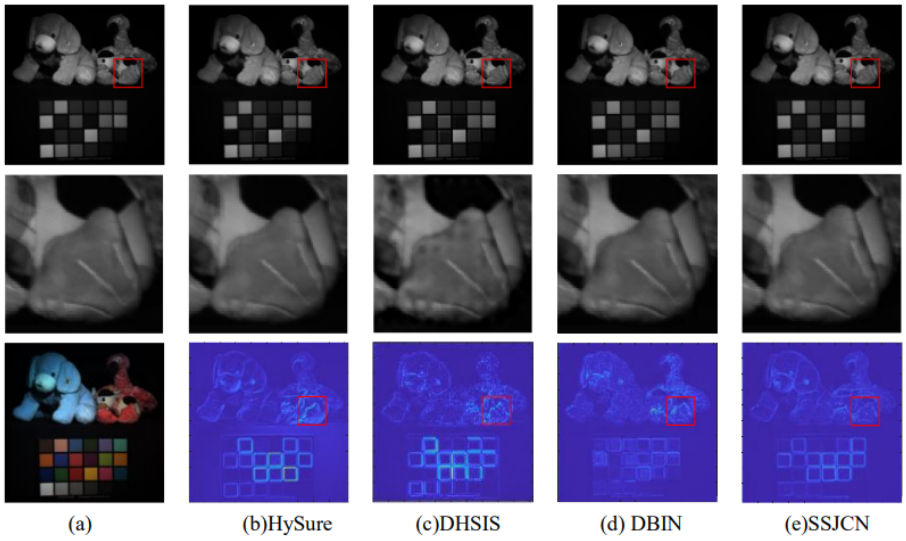

Fig. 4. The results from ***stuffed toys*** at band 20 in the CAVE dataset. (a) the ground truth at band 20 and the HR-MSI. (b-d) reconstructed images and the corresponding error maps after image enhancement while light color represents the error.

4.2 Comparison with Other Methods on CAVE

We take the HSIs from the database as the ground truths. We first blur the ground truths with a Gaussian filter and then down-sample the blurred image by a factor of 1/4. The result of the down-sampling is the simulated LR-HSI. While HR-MSI is generated by multiplying the HS-HSI and the spectral response matrix, and the total number of channels for the HR-MSI is 3. As an image of

size 512∗512 is a heavy burden for reading data with CPU and training with GPU, we take 8∗8 blocks of images from the training set and use the extracted blocks for training.

The test set is being processed in the same way as the training set and the final restored images are obtained by patching the resulting images together in sequence. To better discern the difference, the image enhancement process was implemented on the error maps. The second row in Fig. 4 is a local enlargement of the results obtained by the different methods in the first row, where the results obtained by DHSIS have a clear rectangular block distortion. Although HySure and DBIN maintain the overall structure of the image, obviously results obtained by our proposed method have the least error with the original image. And also, according to the results shown in Table 1, the performance of the proposed method on CAVE dataset was best than those of other methods.

Table 1. Average performance of the compared methods of CAVE dataset.

	PSNR	**SSIM**	**SAM**	**EGRAS**
	$+\infty$	1	0	0
HySure	40.5841	0.9779	6.2523	2.5095
DHSIS	45.1842	0.9903	3.3527	1.3427
DBIN	47.2403	0.9933	3.2230	1.1669
SSJCN	**48.5434**	**0.9937**	**3.0916**	**1.0165**

4.3 Comparison with Other Methods on Harvard

The images in the Harvard data are processed in the same way as the CAVE data. Fig. 5 shows the reconstructed results and the corresponding error maps, again with image enhancement for ease of observation. Combining Fig. 5 and Table 2 we can clearly see that our proposed method yields the lowest error results.

Table 2. Average performance of the compared methods of Harvard dataset.

	PSNR	**SSIM**	**SAM**	**EGRAS**
	$+\infty$	1	0	0
HySure	44.5991	0.9788	3.9709	2.8675
DHSIS	45.7591	0.9812	3.7445	3.1335
DBIN	46.1493	0.9839	3.6503	2.9645
SSJCN	**47.1581**	**0.9847**	**3.3153**	**2.1151**

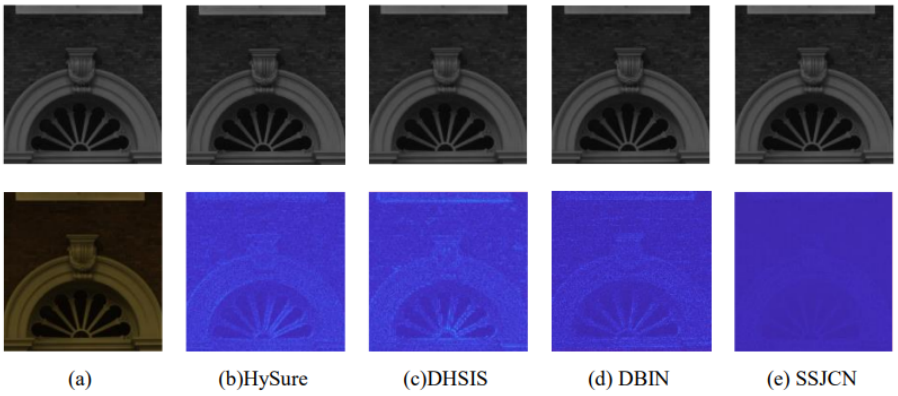

Fig. 5. The results at band 20 of the selected part in the Harvard dataset. (a) the ground truth at band 20 and the HR-MSI. (b-d) reconstructed images and the corresponding error maps after image enhancement while light color represents the error.

4.4 Effectiveness of Degradation Block

As mentioned in the previous section, we hold the opinion that learning the degradation model and error maps to improve the accuracy of the fusion results. Therefore, in this section we demonstrate the effectiveness of the degradation model of the proposed end-to-end model. Take CAVE dataset as an illustration, the results are shown in Fig. 6. Although the output image after RB (1) of Fig. 3 is visually close to the original image, we can see a considerable amount of rectangular deformation in the error image in the Fig. 6, which is reflected in the LR HSI acquired from the degradation of DB (2). It can be inferred that the degradation model in this experiment effectively preserves the details and structural information of the degraded images, which helps to improve the network results. Moreover, we can also observe that the estimation error map outputted from DB (2) is very close to the error between the results reconstructed after RB (1) and the true value, which indicates that the correction map is effective. Besides, performing two iterations on the input data further improves quality assessment values, and the comparison results showed on Table 3.

Table 3. The proposed methods with different iteration numbers of CAVE dataset.

	PSNR	**SSIM**	**SAM**	**EGRAS**
	$+\infty$	1	0	0
SSJCN(1)	47.1570	0.9928	3.2652	1.1584
SSJCN(2)	48.5434	0.9937	3.0916	1.0165

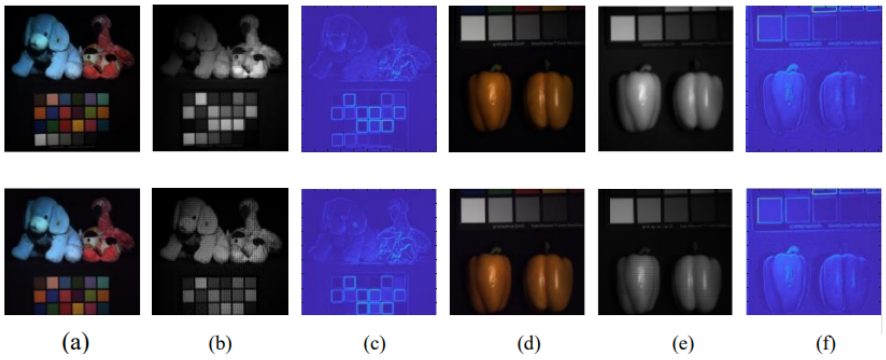

Fig. 6. The results from ***stuffed toys*** and ***real and fake apples*** at band 30. (a, d) first row: original RGB image, second row: the HR MSI obtained after DB (2). (b, e) original LR HSI and the LR HSI obtained after DB (2). (c, f) first row: the error map between the reconstructed result after RB (1) and the true value, second row: the estimated error map in DB2.

5 Conclusion

In this article, a spatial-spectral joint correction network is proposed for HS-MS fusion. SSJCN consists of degradation blocks, error map fusion blocks and the reconstruction blocks, which are used to simulate the degradation mechanism and make corrections to the initialized data respectively. The parameters of network are optimized by minimizing the loss between the outputs and the ground truth. The comparison results between the proposed method and other SOTA methods demonstrate the effectiveness of the proposed method.

References

1. Dian, R., Li, S., Guo, A., Fang, L.: Deep hyperspectral image sharpening. IEEE Transactions on Neural Networks and Learning Systems (2018)
2. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Computer Vision – ECCV 2014 (2014)
3. Dong, W., Fu, F., Shi, G., Cao, X., Wu, J., Li, G., Li, X.: Hyperspectral image super-resolution via non-negative structured sparse representation. IEEE Transactions on Image Processing (2016)
4. Han, X.H., Shi, B., Zheng, Y.: Ssf-cnn: Spatial and spectral fusion with cnn for hyperspectral image super-resolution. In: 2018 25th IEEE International Conference on Image Processing (ICIP) (2018)
5. Han, X.H., Zheng, Y., Chen, Y.W.: Multi-level and multi-scale spatial and spectral fusion cnn for hyperspectral image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops (2019)

6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
7. Kwon, H., Tai, Y.W.: Rgb-guided hyperspectral image upsampling. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
8. Li, S., Dian, R., Fang, L., Bioucas-Dias, J.M.: Fusing hyperspectral and multispectral images via coupled sparse tensor factorization. IEEE Transactions on Image Processing (2018)
9. Liu, P., Xiao, L., Li, T.: A variational pan-sharpening method based on spatial fractional-order geometry and spectral–spatial low-rank priors. IEEE Transactions on Geoscience and Remote Sensing (2018)
10. Loncan, L., de Almeida, L.B., Bioucas-Dias, J.M., Briottet, X., Chanussot, J., Dobigeon, N., Fabre, S., Liao, W., Licciardi, G.A., Simões, M., Tourneret, J.Y., Veganzones, M.A., Vivone, G., Wei, Q., Yokoya, N.: Hyperspectral pansharpening: A review. IEEE Geoscience and Remote Sensing Magazine (2015)
11. Marinelli, D., Bovolo, F., Bruzzone, L.: A novel change detection method for multitemporal hyperspectral images based on binary hyperspectral change vectors. IEEE Transactions on Geoscience and Remote Sensing (2019)
12. Simões, M., Bioucas-Dias, J., Almeida, L.B., Chanussot, J.: A convex formulation for hyperspectral image superresolution via subspace-based regularization. IEEE Transactions on Geoscience and Remote Sensing (2015)
13. V, S., Ealai Rengasari, N.: A survey of hyperspectral image segmentation techniques for multiband reduction. Australian Journal of Basic and Applied Sciences (2019)
14. Wang, W., Zeng, W., Huang, Y., Ding, X., Paisley, J.: Deep blind hyperspectral image fusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
15. Xu, S., Amira, O., Liu, J., Zhang, C.X., Zhang, J., Li, G.: Ham-mfn: Hyperspectral and multispectral image multiscale fusion network with rap loss. IEEE Transactions on Geoscience and Remote Sensing (2020)
16. Yan, H., Zhang, Y., Wei, W., Zhang, L., Li, Y.: Salient object detection in hyperspectral imagery using spectral gradient contrast. In: 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (2016)
17. Zhao, H., Gallo, O., Frosio, I., Kautz, J.: Loss functions for image restoration with neural networks. IEEE Transactions on Computational Imaging (2017)