HAM-MFN: Hyperspectral and Multispectral Image Multiscale Fusion Network With RAP Loss

Shuang Xu, Ouafa Amira, Junmin Liu, *Member, IEEE*, Chun-Xia Zhang, Jiangshe Zhang, and Guanghai Li

***Abstract*— The fusion of hyperspectral image (HSI) and multispectral image (MSI) is one of the most significant topics in remote sensing image processing. Recently, deep learning (DL) has emerged as an important tool for this task. However, existing DL-based methods have two drawbacks, that is, limited ability for feature extraction and suffering from spectral distortion. To address these issues, this article presents a novel neural network, where sophisticated techniques are employed, including network-in-network convolutional unit, batch normalization, and skip connection. To make full use of the MSI, the proposed model fuses HSI and MSI at different scales. Besides, this article presents a new loss function, called RMSE, angle and Laplacian (RAP) loss (the combination of the relative mean squared error, angle loss, and Laplacian loss), to deal with both spatial and spectral distortions. Experiments conducted on four data sets have verified the rationality of network structure and the proposed loss function and demonstrated that the proposed novel model outperforms state-of-the-art counterparts.**

***Index Terms*— Angle loss, convolutional neural network (CNN), hyperspectral image (HSI), image fusion, Laplacian loss, multispectral image (MSI).**

I. Introduction

IN GENERAL, the hyperspectral images (HSIs) can acquire a scene in hundreds of contiguous spectral bands covering the visible and infrared ranges. By this virtue, HSI has boosted several investigations, including unmixing [1] and classification [2]. Although containing enriched spectral information, HSIs suffer from low spatial resolution due to hardware limitations [3]. In contrast, multispectral images (MSIs) are with fewer spectral information but higher spatial resolution. The ideal remote sensing image should be with both high spectral and spatial resolutions. One of the promising techniques is to fuse HSI and MSI to achieve this end.

Manuscript received July 14, 2019; revised October 15, 2019 and November 20, 2019; accepted January 4, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0102201 and Grant 2018YFC0809001, in part by the National Natural Science Foundation of China under Grant 61877049, Grant 11991023, Grant 11671317, and Grant 61976174, and in part by the Fundamental Research Funds for the Central Universities under Grant xzy022019059. *(Corresponding authors: Junmin Liu; Jiangshe Zhang.)*

Shuang Xu, Ouafa Amira, Junmin Liu, Chun-Xia Zhang, and Jiangshe Zhang are with the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: shuangxu@stu.xjtu.edu.cn; amira_ouafa@stu.xjtu.edu.cn; junminliu@mail.xjtu.edu.cn; cxzhang@mail.xjtu.edu.cn; jszhang@mail.xjtu.edu.cn).

Guanghai Li is with the Science and Technology Department, China Special Equipment Inspection and Research Institute, Beijing 100029, China (e-mail: liguagnhai@csei.org.cn).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2020.2964777

This task is closely related to pansharpening, which fuses the MSI and panchromatic image. Since pansharpening often ignores the properties of the HSI, the early exploration of the fusion of HSI and MSI mainly focuses on the adaptation of pansharpening techniques. Thereafter, with the development of machine learning, the research community begins to exploit various tools for a fusion of MSI and HSI, including matrix factorization [4]–[6], sparse representation, and dictionary learning.

Nowadays, deep learning (DL) has made great success in many fields. By the virtue of nonlinear activation and deep hierarchical structure, DL is prone to outperform traditional models. Some DL-based fusion methods have been presented in recent years. However, there are still several problems.

1) Most networks are supervised by mean squared error (MSE) and tend to output blurry images. To the best of our knowledge, there is only one method [7] attempting to deal with this issue.
2) Enriched spectral information is one of the significant differences between HSI and other images. Currently, all networks are devoted to reducing spatial distortion, while neglecting the problem of spectral distortion. Nonetheless, does low spatial distortion necessarily lead to low spectral distortion? It remains a question.
3) The existing methods do not use advanced techniques, such as batch normalization and residual learning.

In this article, to address the above problems, we propose a novel HSI and MSI multiscale fusion network (HAM-MFN). The main contributions are twofold that are as follows.

1) Our first contribution is that we design a novel network structure for HSI and MSI fusion. It consists of two branches separately extracting features of low-resolution HSI (LR-HSI) and MSI. To enhance the capability of feature extraction, the basic convolutional unit is designed in the network-in-network fashion, inspired by GoogleNet [8]. When going deeper, LR-HSI is gradually upscaled and the feature maps of LR-HSI and MSI are fused at different scales. In addition, to boost convergence, there is a skip connection [9] between LR-HSI's interpolation and its final feature map.
2) The second contribution lies in the presentation of a new loss function called RMSE, angle and Laplacian (RAP) loss, in which both spatial and spectral distortions are encoded. Specifically, the low-frequency spatial distortion is described by the relative MSE. Then, inspired by the second derivative of an image, we present the Laplacian loss to measure the high-frequency spatial

textures. At last, spectral distortion is described by a newly proposed angle loss.

According to the experiments conducted on four data sets, our model significantly outperforms state-of-the-art methods. The experiments also verify the rationality of RAP loss.

The rest of this article is organized as follows. Section II reviews the related work on the fusion of the HSI and MSI. Section III presents the new network structure and loss function. Numerical experiments are shown in Section IV. At last, conclusions are drawn in Section V.

II. Related Work

A. Traditional Methods

Most of the traditional methods are based on the following generative model. Let the desired high-resolution HSI (HR-HSI) be denoted by $\mathbf{X} \in \mathbb{R}^{\mathrm{HW}\times B}$, where H, W and B are its height, width, and the number of bands, respectively. The MSI $\mathbf{Y} \in \mathbb{R}^{\mathrm{HW}\times b}$ is obtained by applying a spectral response function $\mathbf{R} \in \mathbb{R}^{B\times b}$ to HR-HSI, while the LR-HSI $\mathbf{Z} \in \mathbb{C}^{hw\times B}$ is obtained by applying a down-sampling matrix $\mathbf{C} \in \mathbb{R}^{hw\times \mathrm{HW}}$. In formula, this generative model is given by

$$\begin{aligned}\mathbf{Y} &= \mathbf{XR} + \epsilon_y \\ \mathbf{Z} &= \mathbf{CX} + \epsilon_z \end{aligned} \tag{1}$$

where ϵ_y and ϵ_z denote the noise.

Matrix factorization can be naturally applied to this problem. Inspired by spectral unmixing, Yokoya *et al.* [4] design a coupled nonnegative matrix factorization (CNMF) method. CNMF estimates endmembers' spectral signatures and high-resolution abundance maps from HR-HSI and MSI, respectively. Integrating them, HR-HSI can be recovered according to the generative model described by (1). Similar to Yokoya *e*t al. [4], Lanaras *et al.* [10] also added several constraints according to the physical properties of image process and spectral unmixing. Considering the local low-rank property of the HSI and MSI, Zhou *et al.* [5] partitioned the HSI and MSI into many patches, and then each pair was fused by CNMF. At last, all fused patches were combined to reconstruct the overall image. Recently, Lin *et al.* [6] presented a convex formulation on this problem by incorporating CNMF with sparsity and sum-of-squared-distances penalties.

Beyond matrix factorization, a number of researchers investigated this problem from a view of probability. For example, Eismann and Hardie [11] reformulated the model described by (1) and make the maximum *a posteriori* estimation. Different from the literature [11], Wei *et al.* [12] imposed the matrix Gaussian distributions on LR-HSI and MSI. By means of maximum likelihood estimation, they cast the original problem into solving a Sylvester equation. Thereafter, in the literature, Wei *et al.* [13] solved this problem in a more robust and efficient way by means of the Woodbury formula. Recently, Lin *et al.* [14] built a hierarchical Bayesian graph for this generative model (1), and it is efficiently solved by variational expectation-maximization.

B. DL-Based Methods

The fusion of HSI and MSI is also closely related to natural image super-resolution, which has recently been excessively investigated via DL methodology [15]–[18]. Inspired by this, research communities started to exploit the convolution neural networks (CNNs) to improve the spatial resolution of the HSI. One of the earliest explorations is the pansharpening neural network (PNN) [19]. Similar to super-resolution CNNs [15], PNNs also consist of three convolutional layers to regress a high-resolution image and their input is the concatenation of the upscaled MSI and panchromatic image along channels. Although the PNN is designed for pansharpening problem, it can be directly applied to the fusion of the HSI and MSI. Thereafter, Palsson *et al.* [20] designed a network that is similar to PNN, but the 3-D-convolutional layers are exploited in the feature extracting stage. Following PNN, Yuan *et al.* [21] proposed the multiscale and multidepth CNN (MSDCNN) to enhance the ability of feature extraction via multiscale convolutional units. However, these methods regard the panchromatic image as a band and concatenate it with MSI as an input of the network, which ignores the distinctive characteristics of the MSI and panchromatic image.

To overcome this drawback, a growing number of researchers decided to abandon the single branch network. Shao and Cai [22] designed a remote sensing image fusion neural network (RSIFNN) which contains two branches separately extracting features of two types of images. Furthermore, an element-wise addition layer is used to fuse two feature maps. In this manner, RSIFNN is able to fully exploit the spectral information of MSI and spatial information of the panchromatic image. Yang *et al.* [23] use 1-D and 2-D convolutional layers to extract spectral and spatial information of HSI and MSI, respectively. Then, the feature maps are merged and reshaped to a vector. At last, the final HR-HSI is recovered by several fully connected layers. Liu *et al.* [24] propose a two-stream fusion network (TFN), where the basic idea is similar to [23], while TFN adopts the fully convolutional architecture. Very recently, Zhou *et al.* [7] propose a pyramid fully CNN (PFCN) to gradually reconstruct HR-HSI in the coarse-to-fine fashion. In the encoder block of PFCN, it extracts spectral information of LR-HSI and obtains a latent image. Then, in the pyramid fusion block, the spatial information of the MSI is integrated with the latent image in the form of a Gaussian pyramid. In the meantime, PFCN employs the gradient difference loss (GDL) to add detail information. At last, it is worth mentioning the deep HSI sharpening model proposed by Dian *et al.* [25], where a deep convolutional network and a model-based approach are combined. However, this model is trained on a part of bands and is tested on the rest of the bands. This manner breaks the integrity of the spectrum.

III. HSI and MSI Fusion Neural Network

In this section, to deal with the drawbacks of CNN-based methods, we present a new network structure and RAP loss. For adapting to the current framework of DL, we use tensors instead of matrices to represent images in the rest of this article. More importantly, both HSI and MSI are essentially tensor cubes, and reshaping them into matrices may reduce the spatial information.

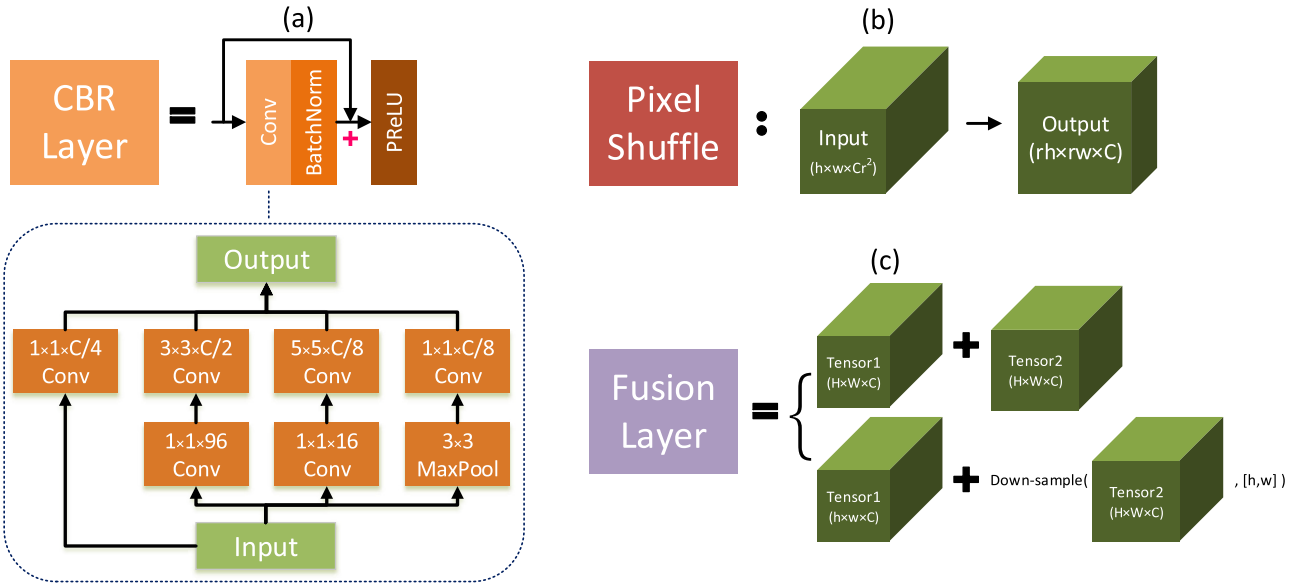


Fig. 1. Basic layers in HAM-MFN. (a) CRB layer. (b) Pixel shuffle. (c) Fusion layer.

A. Basic Layers

Before illustrating the overall structure, we would like to introduce the basic layers of HAM-MFN.

1) CBR Layer: In our network, three units, **c**onvolution, **b**atch normalization, and parametric **R**eLU, consist of the basic feature extracting module, called the CBR layer, as displayed in Fig. 1(a). In order to reduce computation and to enrich the representation extracted by the CBR layer, the convolutional unit is designed in a network-in-network fashion. Given an input, we employ 1×1, 3×3, 5×5 convolutions, and a 3×3 max pooling to extract feature with varied receptive fields. The convolutional unit preserves resolution of input with zero padding. Then, to prevent gradient diffusion, convolutional unit is followed by batch normalization and after that there is a shortcut. At last, we select parametric ReLU as an activation function for flexibility, which is defined by $\mathrm{PReLU}(x) = \max(0, x) + w \min(0, x)$, where w is a learnable parameter. In conclusion, given an input $\mathcal{A}$, the CBR layer can be expressed as the following equation:

$$\mathcal{B} = \mathrm{CBR}(\mathcal{A}) = \mathrm{PReLU}\left(\mathcal{A} + \mathrm{BN}(\mathrm{Conv}(\mathcal{A}))\right). \tag{2}$$

2) Pixel Shuffle Layer: In this layer, the spatial resolution of input tensor is improved. As shown in Fig.1(b), the key idea of pixel shuffle is to reshape input $\mathcal{A} \in \mathbb{R}^{H \times W \times Cr^2}$ into a new tensor $\mathcal{B} \in \mathbb{R}^{rH \times rW \times C}$, where r is the immediate upscale factor (in Fig. 2, $r = 2$). Formally, we have

$$\begin{aligned} \mathcal{B}_{i,j,c} &= \mathrm{PS}(\mathcal{A}) \\ &= \mathcal{A}_{\lfloor i/r \rfloor, \lfloor j/r \rfloor, Cr\mathrm{mod}(j,r)+C\mathrm{mod}(i,r)+c} \end{aligned} \tag{3}$$

where mod is modulo operator and $\lfloor \cdot \rfloor$ denotes rounding a value toward negative infinity. Since this operator leads to reduction of channels, we design a 1×1 convolutional unit after the pixel shuffle layer to keep the number of channels unchanged. The pixel shuffle layer can be replaced by interpolation operators, and it is found that they perform similarly. Note that there are other upscaling operators, such as transposed convolution and partial inverse of pooling. Nonetheless, both of them suffer from checkerboard artifacts [26].

3) Fusion Layer: In this layer, the feature maps of HSI and MSI are fused. As shown in Fig. 1(c), it is actually an element-wise addition operator, that is

$$\mathrm{Fusion}(\mathcal{A}, \mathcal{B})_{ijc} = \mathcal{A}_{ijc} + \mathcal{B}_{ijc}. \tag{4}$$

If two input tensors are not in the same shape, we, at first, rescale a tensor with desired shape and then make element-wise addition.

B. Network Design

The overall architecture is displayed in Fig. 2. It contains two branches separately extracting the feature maps of LR-HSI and MSI. As the network goes deeper, LR-HSI is gradually upscaled and feature maps are fused at different scales.

In block 1, the network is fed with MSI $\mathcal{Y} \in \mathbb{R}^{H \times W \times b}$ and LR-HSI $\mathcal{Z} \in \mathbb{R}^{h \times w \times B}$. It is assumed that the spatial size of MSI is four times larger than that of the HSI, that is, $(H, W) = 4(h, w)$. Since the number of channels for two input images are different, we utilize two initial CBR layers to make their feature maps to be with the same number of channels (denoted by C) for convenience of following operations. Then, the two CBR layers are applied, and their output feature maps are fused at scale (h, w). As stated above, the spatial size of MSI's feature map is larger than that of the HSI's feature map. So to be compatible with HSI, the feature map of the MSI is downsampled by interpolation, in the fusion layer. The outputs of block 1 are denoted by MS1 and HS1, which, in formula, can be expressed by

$$\begin{aligned} \mathrm{MS1} &= \mathrm{CBR}_1^{\mathrm{MS}}\left(\mathrm{CBR}_0^{\mathrm{MS}}(\mathcal{Y})\right) \\ \mathrm{HS1} &= \mathrm{Fusion}\left(\mathrm{CBR}_1^{\mathrm{HS}}\left(\mathrm{CBR}_0^{\mathrm{HS}}(\mathcal{Z})\right), \mathrm{MS1}\right). \end{aligned} \tag{5}$$

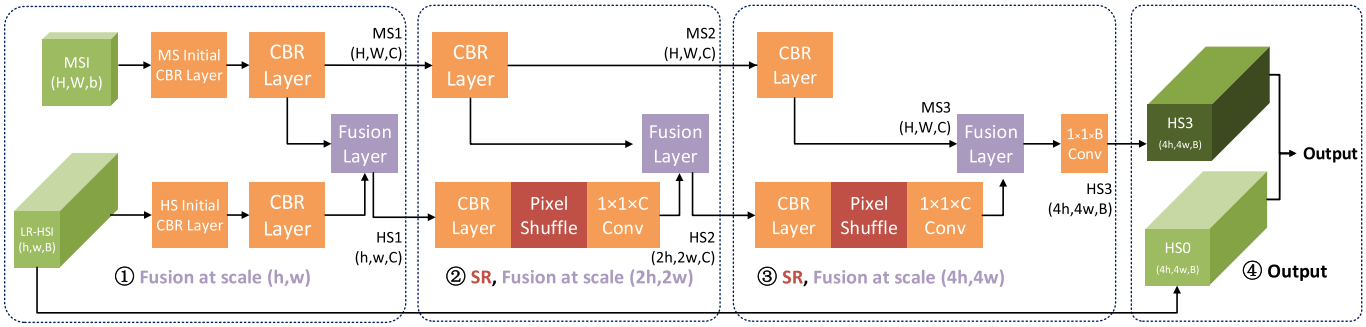


Fig. 2. Structure of HAM-MFN. The network upscales LR-HSI four times, that is, $(H, W) = 4(h, w)$.

In block 2, for the MSI's feature map (MS1), we still utilize a CBR layer to extract feature. While for LR-HSI's feature map (HS1), we employ CBR, pixel shuffle, and $1 \times 1 \times C$ convolutional layers to extract feature and upscale the spatial resolution with shape $(2h, 2w)$. At last, two immediate feature maps are fused at scale $(2h, 2w)$. The operations of block 2 can be expressed by

$$\begin{aligned} \text{MS2} &= \text{CBR}_2^{\text{MS}}(\text{MS1}) \\ \text{HS2} &= \text{Fusion}\big(\text{Conv}\big(\text{PS}\big(\text{CBR}_2^{\text{HS}}(\text{HS1})\big)\big), \text{MS2}\big). \end{aligned} \tag{6}$$

In block 3, we repeat the same operations in block 2 and implement fusion at scale $(4h, 4w)$. Since it is ready to recover HR-HSI, a $1 \times 1 \times B$ convolutional unit is applied to HSI's feature map so as to match the number of channels. In formula, there are

$$\begin{aligned} \text{MS3} &= \text{CBR}_3^{\text{MS}}(\text{MS2}) \\ \widetilde{\text{HS3}} &= \text{Fusion}\big(\text{Conv}\big(\text{PS}\big(\text{CBR}_2^{\text{HS}}(\text{HS2})\big)\big), \text{MS3}\big) \\ \text{HS3} &= \text{Conv}(\widetilde{\text{HS3}}). \end{aligned} \tag{7}$$

In the last block, we reconstruct the output (HR-HSI). Since there is a high degree of similarity between HR and LR images, a skip connection is created between the upscaled image, HR-HS0, and final feature map, HS3, to prevent from learning redundant information.

Remark 1: The reasons for designing the multiscale fusion network is twofold. First, it is reported that it is difficult to train a single scale network if the scale factor is large, and, as a substitute, progressively reconstructing the high-resolution images in a coarse-to-fine manner is a better choice [17]. Second, it has been pointed out that the single-scale network leads to a high computation cost [18], [27]. The multiscale network, to some degree, saves the cost since it gradually increases the spatial resolution. Last but not least, excessive references have proved that shallower and deeper layers extract low-level spatial-visual and high-level semantic information, respectively. Only fusing the feature map of a specific layer is inadequate to recover the high-resolution images.

Remark 2: In the following experiments, we set feature channel C to 256. Although the architecture displayed in Fig. 2 is designed for four times upscaling, we can slightly modify the network architecture so as to get desired upscale factors, such as add/remove blocks and change immediate upscale factor r in the pixel shuffle layer. For example, removing block 3 or setting r to 3 in block 3 corresponds to two and six times upscaling, respectively.

C. Loss Function

We present a new loss function, RAP, for HSI and MSI fusion problem. RAP loss consists of three parts, that is

$$L = \text{RMSE} + \lambda_1 L_{\text{angle}} + \lambda_2 L_{\text{Lap}}. \tag{8}$$

Relative MSE (RMSE) and Laplacian loss measure the differences of low- and high-frequency textures, and angle loss describes the difference of spectrum. The hyper-parameters λ_1 and λ_2 are nonnegative weight coefficients.

1) RMSE: Instead of MSE, we use the RMSE to measure the difference between output and groundtruth in image space. In formula, RMSE is defined by

$$\text{RMSE}(\mathcal{X}, \hat{\mathcal{X}}) = \sqrt{\frac{1}{B}\sum_{b=1}^{B} \frac{\sum_{i,j}(x_{ijb} - \hat{x}_{ijb})^2}{\sum_{i,j} x_{ijb}^2}} \tag{9}$$

where x_{ijb} and $\hat{x}_{ijb}$ are the values of (i, j) pixel in bth band for groundtruth and output, respectively.

2) Laplacian Loss: As is well-known, RMSE tends to result in oversmoothed images and does not necessarily guarantee the consistency of high-frequency textures. Recently, researchers have been devoted to making the results satisfy human perception. In general, there are two ways to achieve this end. The first one is to use perceptual loss measuring the images' differences in feature space. Specifically, high-level image feature representations are extracted by a pretrained CNN [28], such as Visual Geometry Group Net [29]. The other way is adversarial training [30]. That is, we regard super-resolution mapping as a generation network, whose output is used to fool the discriminator network. In this manner, adversarial training encourages the generator to learn the manifold of groundtruth images. Although both approaches recover finer texture details for natural images, we emphasize that they cannot be applied to our task. The reason is that both perceptual loss and adversarial training generate artifacts making output satisfy human perception (refer to [28, Fig. 8] and [30, Fig. 2]). Nonetheless, the artifacts harm the HSI super-solution task which requires an accurate output, since the purpose of this task is to analyze HSI rather than view and admire it. For example, the leaf area index [31] can be calculated in the HSI to evaluate seagrass

density, whereas the artifacts definitely reduce the accuracy. As an alternative, we propose the Laplacian loss defined by

$$L_{\text{Lap}} = \text{RMSE}(\text{LapConv}(\mathcal{X}), \text{LapConv}(\hat{\mathcal{X}})) \tag{10}$$

where LapConv denotes the Laplacian convolution whose kernel matrix is

$$f = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}. \tag{11}$$

Actually, Laplacian convolution corresponds to the second spatial derivative of an image

$$\begin{aligned} \text{LapConv}(\mathbf{I}) &= \frac{\partial^2 \mathbf{I}}{\partial x^2} + \frac{\partial^2 \mathbf{I}}{\partial y^2} \\ &= \mathbf{I}(x+1, y) + \mathbf{I}(x-1, y) + \mathbf{I}(x, y+1) \\ &\quad + \mathbf{I}(x, y-1) - 4\mathbf{I}(x, y). \end{aligned} \tag{12}$$

This operator is sensitive to edges of an image and has been widely applied to edge detection and image sharpening. Therefore, Laplacian loss can measure difference of two images' high-frequency textures. It is worth pointing out the distinction between Laplacian loss and GDL that is proposed in [7]. Similar to Laplacian loss, GDL uses the first-order gradient to extract high-frequency textures. The horizontal and vertical gradients for a gray scale image are defined by

$$\Delta \mathbf{I}^h_{ij} = |\mathbf{I}_{ij} - \mathbf{I}_{i-1,j}|, \text{ and } \Delta \mathbf{I}^v_{ij} = |\mathbf{I}_{i,j-1} - \mathbf{I}_{ij}| \tag{13}$$

respectively. Based on them, GDL is expressed by

$$L_{\text{gdl}} = \sum_{i,j} |\Delta \mathbf{I}^h_{ij} - \Delta \hat{\mathbf{I}}^h_{ij}| + |\Delta \mathbf{I}^v_{ij} - \Delta \hat{\mathbf{I}}^v_{ij}|. \tag{14}$$

Equation (14) is defined for the gray scale image (that is, with one channel). For the HSI, GDL is applied to each band, and all the outputs are combined.

3) Angle Loss: Although RMSE and Laplacian loss can guide networks to reduce spatial distortion, they cannot meet the demand of our task. The reason lies in that, different from natural images with three bands, HSI collects hundreds of contiguous bands providing rich details of the spectral signature of various materials. The loss function should take spectral distortion into account. So, beyond RMSE and Laplacian loss, we present angle loss to measure spectral distortion, which is defined by

$$L_{\text{angle}} = \frac{1}{\text{HW}} \sum_{i,j} \arccos\left(\frac{\hat{\mathcal{X}}_{ij}^T \mathcal{X}_{ij}}{\|\hat{\mathcal{X}}_{ij}\|_2^2 \|\mathcal{X}_{ij}\|_2^2} \right) \tag{15}$$

where $\mathcal{X}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijb})^T$ denotes the values of (i, j) pixel along all bands.

At last, it is worth pointing out that plenty of experiments show that the values of the three loss functions are with a similar magnitude. Therefore, we set $\lambda_1 = \lambda_2 = 1$.

D. Gradient of RAP Loss

Here, we analyze the backpropagation rules for the computation of gradients. According to chain rule, we have

$$\frac{\partial L}{\partial w} = \sum_{i,j,b} \frac{\partial \hat{x}_{ijb}}{\partial w} \left(\frac{\partial \text{RMSE}}{\partial \hat{x}_{ijb}} + \lambda_1 \frac{\partial L_{\text{angle}}}{\partial \hat{x}_{ijb}} + \lambda_2 \frac{\partial L_{\text{Lap}}}{\partial \hat{x}_{ijb}} \right). \tag{16}$$

Recall that $\hat{x}_{ijb}$ is the (i, j) pixel in $\hat{\mathcal{X}}^b$, the bth band of $\hat{\mathcal{X}}$. We mainly focus on three items within brackets in (16). For the first item, we have

$$\frac{\partial \text{RMSE}}{\partial \hat{x}_{ijb}} = \left(\sum_{k=1}^{B} \frac{\|\hat{\mathcal{X}}^k - \mathcal{X}^k\|_2^2}{\|\mathcal{X}^k\|_2^2} \right)^{-\frac{1}{2}} \frac{\hat{x}_{ijb} - x_{ijb}}{\sqrt{B}\|\hat{\mathcal{X}}^b\|_2^2}. \tag{17}$$

As for the second item, we have

$$\frac{\partial L_{\text{angle}}}{\partial \hat{x}_{ijb}} = \frac{1}{\text{HW}} \frac{\partial \arccos\left(\frac{\hat{\mathcal{X}}_{ij}^T \mathcal{X}_{ij}}{\|\hat{\mathcal{X}}_{ij}\|_2 \|\mathcal{X}_{ij}\|_2} \right)}{\partial \hat{x}_{ijb}}. \tag{18}$$

For convenience, let

$$\begin{aligned} \alpha &= \hat{\mathcal{X}}_{ij}^T \mathcal{X}_{ij} = \sum_{k=1}^{B} \hat{x}_{ijk} x_{ijk} \\ \beta &= \|\hat{\mathcal{X}}_{ij}\|_2^2 = \sum_{k=1}^{B} \hat{x}_{ijb}^2, \quad \gamma = \|\mathcal{X}_{ij}\|_2^2 = \sum_{k=1}^{B} x_{ijb}^2. \end{aligned}$$

It is easy to rewrite the second item as

$$\frac{\partial L_{\text{angle}}}{\partial \hat{x}_{ijb}} = \frac{1}{\text{HW}\sqrt{1 - \frac{\alpha}{\beta\gamma}}} \frac{\alpha\beta' - \beta\alpha'}{\beta^2 \gamma} \tag{19}$$

where

$$\alpha' = x_{ijb}, \quad \beta' = \hat{x}_{ijb}. \tag{20}$$

For the last item, recall the definition Laplacian convolution, $\text{LapConv}(\hat{\mathcal{X}}) = \text{cat}\{f * \hat{\mathcal{X}}^k\}_{k=1}^K$, where cat is the concatenating operator for a set of tensors along channels, and $*$ denotes the convolution operator. At the same time, the Laplacian loss can be rewritten as

$$L_{\text{Lap}} = \sqrt{\frac{1}{B} \sum_{k=1}^{B} \frac{\|f * (\hat{\mathcal{X}}^k - \mathcal{X}^k)\|_2^2}{\|f * \mathcal{X}^k\|_2^2}}. \tag{21}$$

Now, it is ready to obtain the gradient of the Laplacian loss, namely

$$\frac{\partial L_{\text{Lap}}}{\partial \hat{x}_{ijb}} = \left(\sum_{k=1}^{B} \frac{\|f * (\hat{\mathcal{X}}^k - \mathcal{X}^k)\|_2^2}{\|f * \mathcal{X}^k\|_2^2} \right)^{-1/2} \frac{\xi_{ij}}{\sqrt{B}\|f * \mathcal{X}^b\|_2^2} \tag{22}$$

where ξ_{ij} is the (i, j) entry of the following matrix:

$$\xi = f^T * (f * (\hat{\mathcal{X}}^b - \mathcal{X}^b)). \tag{23}$$

Plug (17), (19), and (22) into (16), and we could obtain the gradient of RAP loss.

IV. Experiments

In this section, we conducted a series of experiments to study the behavior of our network.

TABLE I
INFORMATION OF DATA SETS

Name	Training image		Testing image	
	Pixel size	Left top position	Pixel size	Left top position
Botswana	400×256	$(1, 1)$	256×256	$(401, 1)$
PU	340×340	$(1, 1)$	260×340	$(351, 1)$
WDC	280×280	$(1001, 1)$	300×280	$(1, 1)$
Urban	128×307	$(1, 1)$	160×160	$(131, 1)$

A. Implementation Details

1) Data Sets: In the experiments, four data sets are employed.

1) *Botswana:* The NASA EO-1 satellite acquired a sequence of data over the Okavango Delta, Botswana in 2001–2004. The Hyperion sensor on EO-1 acquires data at 30-m pixel resolution over a 7.7-km strip in 242 bands covering the 400–2500-nm portion of the spectrum in 10-nm windows. The corrupted bands are removed, and there remain 145 bands, each of which is with 1476×256 pixels.
2) *Pavia University (PU):* It was acquired by the ROSIS sensor during a flight campaign over Pavia, northern Italy. The total field of view is $\pm 8^\circ$ and this sensor covers the spectral range from 430 to 860 nm. There are 103 bands, each of which contains 610×340 pixels.
3) *Washington DC (WDC) Mall:* It was obtained from the hyperspectral digital imagery collection experiment HYDICE) from an airborne hyperspectral data flight line over the Washington, DC urban area. Two hundred and ten bands were collected in the 0.4–2.4-μm region of the visible and infrared spectrum. Some water absorption channels were discarded, resulting in 191 bands, each of which is with 1280×307 pixels.
4) *Urban:* The HYDICE Urban data set was captured over Copperas Cove, near Fort Hood, TX, USA, in October 1995. There are 210 bands in the original data. The bands 1–4, 76, 87, 101–111, 136–153, and 198–210 were removed as noisy or water-absorption bands, leaving 162 bands with a size of 307×307. The spectral and spatial resolutions are 10 nm and 2 m, respectively.

2) Training Details: Our network is trained on paired data set $\{\mathcal{X}_i, (\mathcal{Y}_i, \mathcal{Z}_i)\}$, where $\mathcal{X}_i$,$\mathcal{Y}_i$ and $\mathcal{Z}_i$ denote the HR-HSI, MSI, and LR-HSI, respectively. Nonetheless, the groundtruth, $\mathcal{X}_i$, is often unavailable. Therefore, the Wald protocol [32] is exploited to generate the training data. For our experiments, as shown in Table I, we select two nonoverlapped areas for training and testing, respectively. In the training phase, the training area of the original HSI is cropped into $8R \times 8R$ patches, each of which is regarded as HR-HSI $\mathcal{X}_i$ (that is, the groundtruth). Note R denotes the scale factor here. The MSI $\mathcal{Y}_i$ is obtained by applying a randomly generated spectral response function to HR-HSI $\mathcal{X}_i$. At last, the LR-HSI $\mathcal{Z}_i$ is generated by downsampling HR-HSI $\mathcal{X}_i$ by means of bicubic interpolation. In the following experiment, we set scale factor R to 2, 4, and 8, so the spatial sizes of HR-HSI $\mathcal{X}_i$ are 16×16, 32×32, and 64×64, respectively. Due to the limited space, only the results obtained with $R = 4$ are displayed in Section IV-B, and the others can be referred to in the Supporting Information.

We use Pytorch to implement and train HAM-MFN on a computer with Intel Core i7-8750H CPU at 2.20 GHz and RTX 2080ti GPU. HAM-MFN is optimized by Adam [33] with a learning rate of 1e−3 from 1st to 200th epochs and SGD with a learning rate of 5e−4 from 201st to 250th epochs. In each epoch, there are 1024 pairs of training data $\{\mathcal{X}_i, (\mathcal{Y}_i, \mathcal{Z}_i)\}$. Therefore, there are $1024 \times 250 = 256\,000$ pairs of training samples.

B. Performance Comparison

Several state-of-the-art counterparts are used for comparison, that is, RSIFNN [22], MSDCNN [21], PFCN [7], and TFN [24]. We write the codes by Pytorch and train them over 500 epochs. Except for the DL-based methods, four traditional models are also selected for comparison, that is, CNMF [4], FUSE [12], ICCV15 [10], and MAPSMM [11]. They are implemented with official MATLAB codes. We use peak signal-to-noise ratio (PSNR), spectral angle mapper (SAM), erreur relative globale adimensionnelle de synthèse (ERGAS) and structure similarity (SSIM) for performance evaluation. A fused image is better, if it is with higher PSNR and SSIM, and lower SAM and ERGAS.

Table II reports the quantitative results of all methods on four data sets. As shown in this table, not surprisingly, MSDCNN performs worst in all cases. The main reason lies in the fact that MSDCNN concatenates MSI and LR-HSI along the channels as input and utilizes single branch extract features. In this manner, MSDCNN not only neglects the distinctive characteristics of MSI and LR-HSI, but also employs an improper fusion strategy. In general, RSIFNN does a better job than MSDCNN. The main difference between MSDCNN and RSIFNN is that the latter one employs two CNN branches to separately learn abstract features of MSI and LR-HSI. It is shown that all metrics are significantly improved. The other reason why MSDCNN and RSIFNN perform badly is that they require training with a large epoch number. PFCN and TFN are two typical networks with two branches. It is shown that they tend to achieve better results compared with four traditional methods. Our proposed model, by and large, outperforms others with regard to all metrics, although it is trained with 250 epochs (half of the epoch number of other methods).

The visual inspection shown in Figs. 3–6 also demonstrates the superiority of HAM-MFN. It is reported that MSDCNN and RSIFNN produce blurry images. There is color distortion and noise in images fused by FUSE. It is worth noting that the images produced by PFCN, TFCN, CNMF, ICCV15, MAPSMM, and HAM-MFN are clear, but HAM-MFN does a better job in some local areas. For example, the areas bounded by red rectangles in Figs. 3, 5, and 6, indicate PFCN, ICCV15, and MAPSMM lead to the color distortion. Additionally, they produce blurry areas on WDC data set, as shown in Fig. 4. Intending to get more insights, we plot residual maps, the pixel

TABLE II

PERFORMANCE ON FOUR DATA SETS. THE BEST AND SECOND BEST RESULTS ARE HIGHLIGHTED BY BOLD AND ITALIC TYPEFACES, RESPECTIVELY

Botswana	PSNR	SAM	ERGAS	SSIM
HAM-FNN	**38.1412**	**1.9589**	**1.5959**	**0.9682**
RSIFNN	27.4089	3.1861	4.0627	0.6715
MSDCNN	26.0382	5.9563	5.1780	0.6530
PFCN	31.1278	4.2083	2.7734	0.9069
TFN	*37.3259*	*2.1418*	*1.7818*	*0.9643*
CNMF	33.2778	*2.5860*	2.0843	0.9411
FUSE	29.4583	4.3288	3.4254	0.8527
ICCV15	34.0452	2.5906	2.1079	0.9457
MAPSMM	30.8616	2.8261	2.7959	0.8969

WDC	PSNR	SAM	ERGAS	SSIM
HAM-FNN	**34.0858**	**2.5441**	**2.0808**	**0.9783**
RSIFNN	21.9297	6.5384	8.3966	0.7273
MSDCNN	19.3426	16.9952	10.9418	0.6136
PFCN	26.4934	7.2358	4.7873	0.9257
TFN	*32.7950*	*3.0877*	*2.3830*	*0.9710*
CNMF	29.0670	5.1820	3.4596	0.9402
FUSE	24.0700	8.2218	6.0367	0.8642
ICCV15	27.9971	4.9720	4.0728	0.9385
MAPSMM	25.7453	6.6100	5.1545	0.9034

PU	PSNR	SAM	ERGAS	SSIM
HAM-FNN	**40.8632**	**2.5308**	**1.8052**	**0.9776**
RSIFNN	25.6643	6.0198	8.4957	0.7000
MSDCNN	25.7682	9.2626	8.5494	0.6789
PFCN	39.9351	3.0421	2.0957	0.9724
TFN	*40.6943*	*2.6396*	*1.8221*	*0.9754*
CNMF	32.0136	4.4615	4.1581	0.9332
FUSE	28.9074	8.2179	5.9420	0.8729
ICCV15	32.6323	4.4328	3.9975	0.9359
MAPSMM	30.5987	6.1352	4.8372	0.9050

Urban	PSNR	SAM	ERGAS	SSIM
HAM-FNN	**33.1616**	**3.7129**	**3.2120**	**0.9709**
RSIFNN	21.5555	8.7249	9.8615	0.6321
MSDCNN	19.8254	14.6843	11.8206	0.5002
PFCN	29.6500	6.0412	5.0407	0.9543
TFN	*32.8693*	*3.8801*	*3.5527*	*0.9703*
CNMF	24.7231	10.1844	6.9528	0.8880
FUSE	20.8779	19.9514	11.2009	0.7290
ICCV15	25.3047	8.4236	6.7973	0.8752
MAPSMM	24.0531	12.1368	7.5048	0.8684

of which is defined by $\mathbf{I}_{ij}^{\text{res}} = \sum_{b=1}^{B} |x_{ijb} - \hat{x}_{ijb}|/B$. Darker the residual map is, better the method is. As shown in Fig. 7, we take the Botswana data set as an example (more results are given in the Supporting Information). It is found that HAM-MFN's residual map is darkest, while the evident textures or noise can be observed in other residual maps. This fact indicates that HAM-MFN is a more efficient model to extract high-frequency textures.

Besides spatial visual inspection, we want to obtain more insights into the spectrum. We randomly select a pixel for each data set and display their spectral signatures among different methods in Fig. 8. Apparently, MSDCNN, RSIFNN, PFCN, and FUSE make vast distortions. CNMF, ICCV15, MAPSMM, and TFN, to some extent, avoid spectral distortion, but their spectrum curves cannot meet the demand. For example, their curves from band 50 to 100 in WDC data set severely differ from groundtruth (colored in dark blue). On the whole, HAM-MFN produces the best result. Particularly, in some cases, the curves of HAM-MFN (colored in light blue) are almost coincident with groundtruth. In the meanwhile, we plot the angle maps, each pixel of which is defined by

$$\mathbf{I}_{ij}^{\text{angle}} = \arccos\left(\frac{\sum_{b=1}^{B} x_{ijb}\hat{x}_{ijb}}{\sqrt{\sum_{b=1}^{B} x_{ijb}^2 \sum_{b=1}^{B} \hat{x}_{ijb}^2}}\right).$$

Smaller the angle is, more similar the two spectrum curves are. As displayed in the second row of Fig. 7, HAM-MFN is the best method to avoid spectral distortion.

TABLE III

EVALUATION METRICS OF DIFFERENT LOSS FUNCTIONS

No.	Loss function	PSNR	SAM	ERGAS	SSIM
(a)	RMSE	37.0408	2.0662	1.7817	0.9652
(b)	RMSE+GDL	37.1156	2.0286	1.8331	0.9678
(c)	RMSE+Lap	37.5326	2.1305	*1.6243*	0.9638
(d)	RMSE+Angle	37.0450	**1.9146**	2.2000	0.9669
(e)	RMSE+Angle+GDL	37.0975	*1.9467*	2.0148	*0.9680*
(f)	RAP	**38.1412**	1.9589	**1.5959**	**0.9682**

C. Study on Loss Function

In Section IV-B, the exhaustive experiments have proved the effectiveness and superiority of HAM-MFN. In this section, we study the performance of HAM-MFN with different loss functions on Botswana data set at a scale factor of four, aiming to analyze how each part of the RAP loss contributes to the final performance of our network. At the same time, we are also interested in comparing the performance of GDL and Laplacian loss.

In experiment (a), RMSE is regarded as a baseline loss function, and from experiments (b) to (f), we gradually add different regularizers. The results are reported in Table III. It is shown that if only RMSE is combined with a regularizer, the performance would be improved to some extent. At first, we study GDL and Laplacian loss. In experiments (b) and (c), the loss functions are RMSE plus GDL and Laplacian loss, respectively. It is reported that Laplacian loss is good at recovering the spatial structure and image textures since our network in experiment (c) achieves much higher PSNR and lower ERGAS values. However, GDL is good at dealing with spectral distortion and our network achieves slightly lower SAM value in experiment (b). In experiments (e) and (f), a similar conclusion can be drawn. On the whole, Laplacian loss strikes the balance between spatial and spectral distortion, while GDL does a bad job in spatial structure. Secondly, we find that it is necessary to combine angle loss. For example, the difference between experiments (b)/(c) and (e)/(f) is whether angle loss is included. The metrics reveal that angle loss not only reduces spectral distortion but also significantly decreases spatial deformation. In conclusion, this experiment demonstrates the rationality of RAP loss.

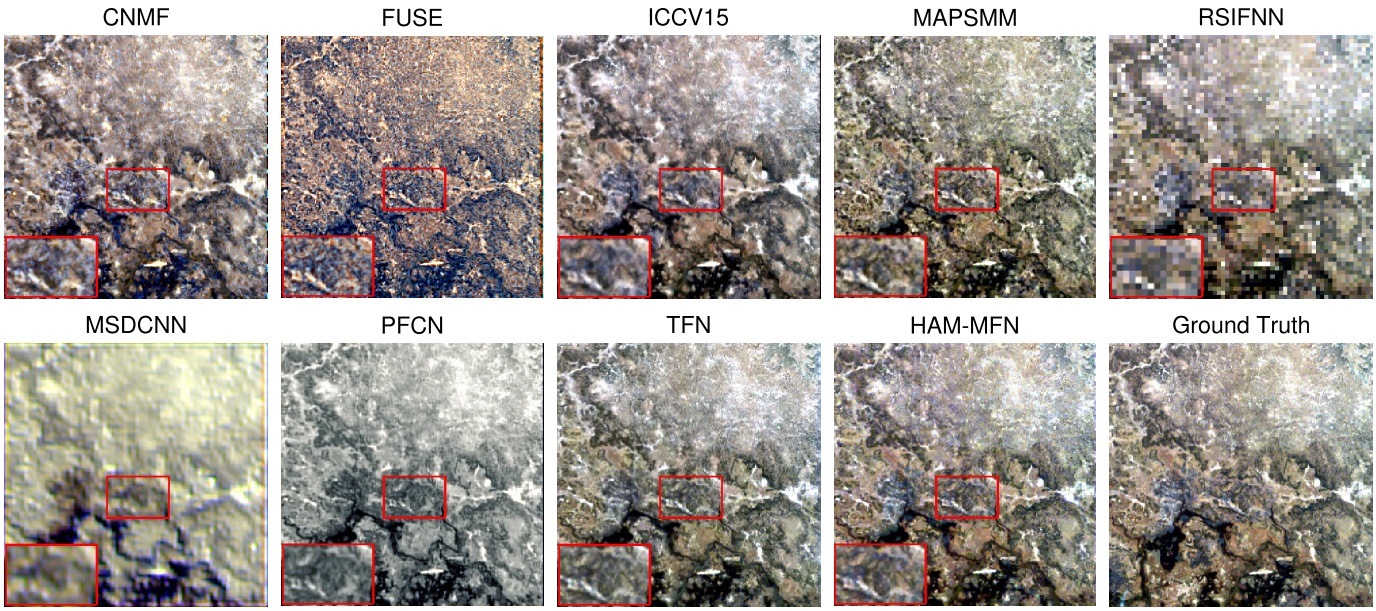


Fig. 3. Fusion results on Botswana. R-G-B channels correspond to band 41-92-135.

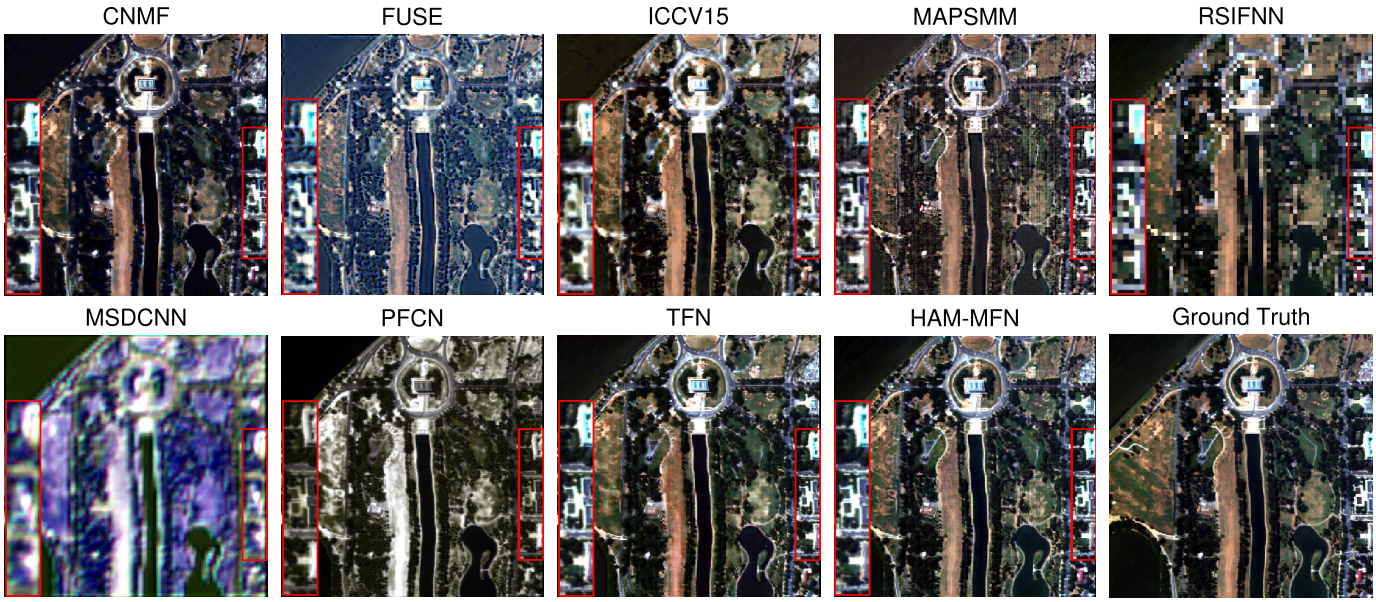


Fig. 4. Fusion results on WDC. R-G-B channels correspond to band 24-40-28.

D. Experiments on a Real World Data Set

In this section, we test methods on a real-world data set. Because there are few publicly available pairs of the HSI and MSI that captured with the same satellite, we employ the MSI and the natural color image of the Roman Colosseum acquired by World View-2. In this experiment, the original MSI of size $419 \times 658 \times 8$ is regarded as LR-HSI, while the natural color image of size $1676 \times 2632 \times 3$ is regarded as MSI. The top half area of LR-HSI ($209 \times 658 \times 8$) and MSI ($836 \times 2632 \times 3$) are used to train networks. Similar to simulated experiments, the Wald protocol is used to generate a training data set. At first, the MSI is downsampled four times. Then, in each iteration, we randomly crop a $32 \times 32 \times 3$ MSI patch as $\mathcal{Y}_i$ and a $32 \times 32 \times 8$ LR-HSI patch as $\mathcal{X}_i$. In the meanwhile, $\mathcal{Z}_i$ is generated by downsampling the LR-HSI patch four times. All networks are trained by Adam with a learning rate of 1e$-$3 over 2000 epochs, and in each epoch, there are 1024 pairs of training data $\{\mathcal{X}_i, (\mathcal{Y}_i, \mathcal{Z}_i)\}$.

In the testing phase, we randomly crop 30 patches in the bottom half to study the performance. The classical methods, FUSE and ICCV15, are not employed, since they raise the numerical error in this experiment. To evaluate the methods, we exploit the multivariate Gaussian (MVG)-based no-reference HSI quality assessment metric proposed in [34], which integrates a group of spectral–spatial quality-sensitive features. The smaller the MVG score is, the better the fused

Fig. 5. Fusion results on PU. R-G-B channels correspond to band 35-98-60.

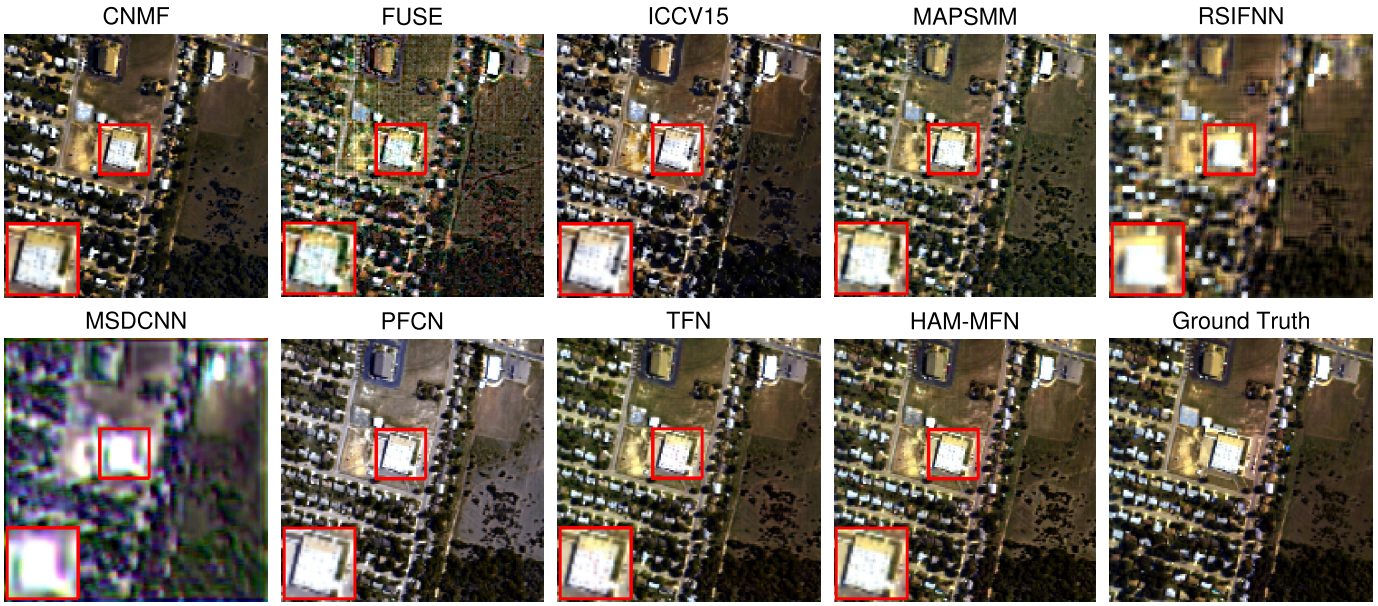


Fig. 6. Fusion results on Urban. R-G-B channels correspond to band 94-116-145.

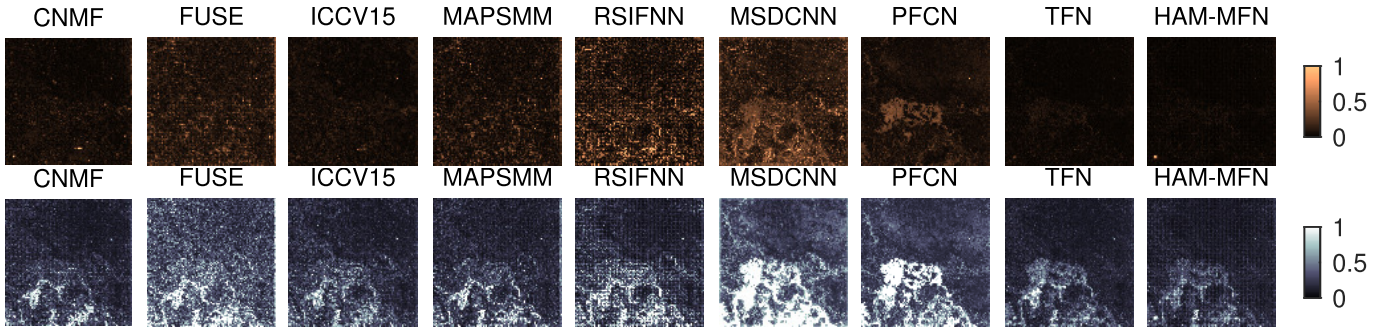


Fig. 7. (Top) Residual and (Bottom) angle maps of Botswana. The pixels in residual map are amplified eight times for ease of visual inspection.

image is. Table IV reports the MVG scores averaged over 30 patches, and it is found that HAM-MFN takes the best place. In addition, Fig. 9 shows the fusion images of three patches (refer to the Supporting Information for more results). The visual inspection evidently demonstrates that the image fused by HAM-MFN is clearer and sharper.

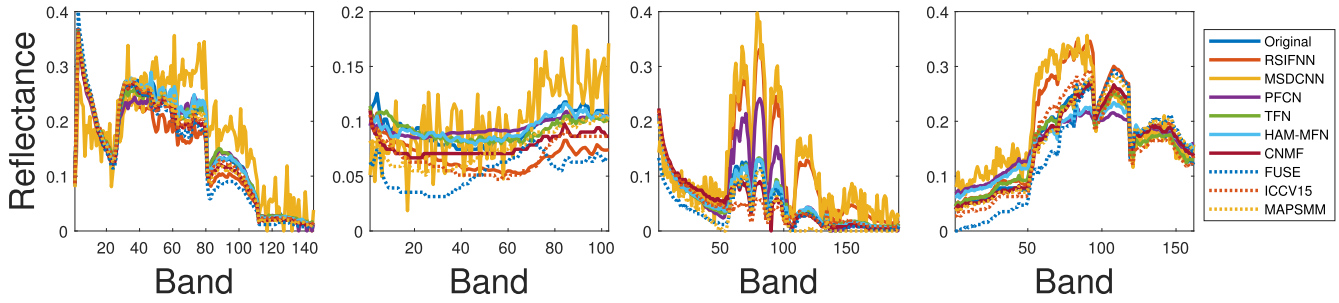


Fig. 8. Spectrum signatures.

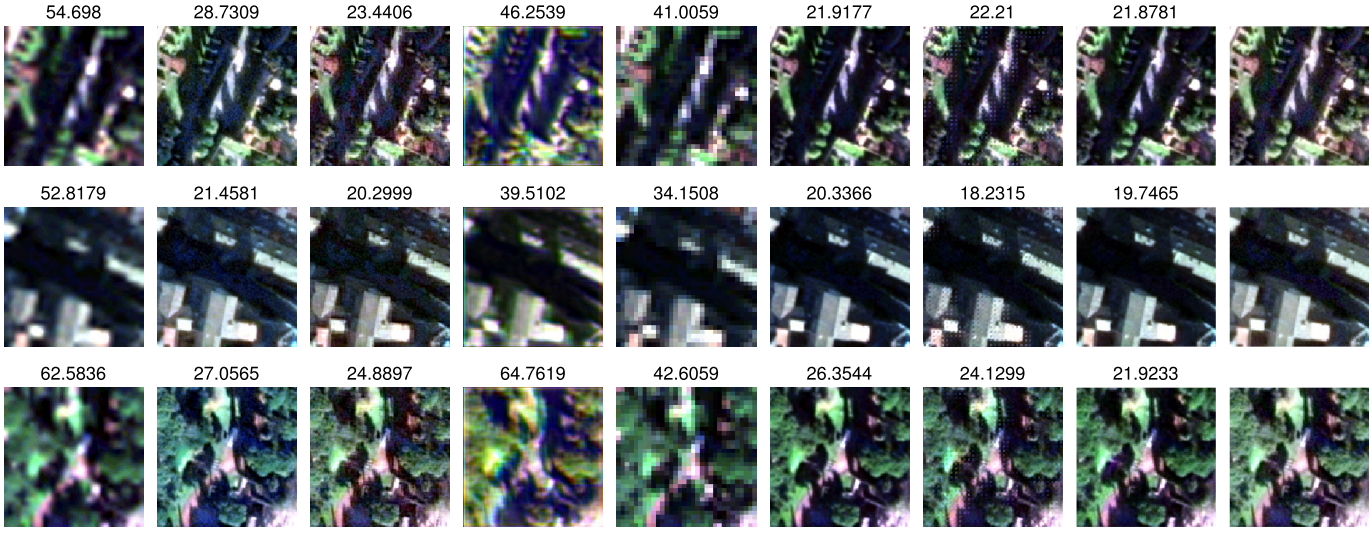


Fig. 9. Fusion results on WV2. R-G-B channels correspond to band 5-3-2. (From left to right) Methods are bicubic, CNMF, MAPSMM, MSDCNN, RSIFNN, PFCN, TFN, HAM-MFN, and MSI. Last column: corresponding MSI patch. The title of each subimage reports its MVG score.

TABLE IV
MEAN MVG SCORE ON WV2 DATA SET. THE BEST AND SECOND BEST RESULTS ARE HIGHLIGHTED BY BOLD AND ITALIC TYPEFACE, RESPECTIVELY

	Bicubic	CNMF	MAPSMM	MSDCNN
Mean	52.5677	27.1646	24.6867	51.4586
	RSIFNN	PFCN	TFN	HAM-MFN
Mean	40.0997	23.8507	*22.7701*	**22.4978**

V. CONCLUSION

In this article, we propose a novel CNN for HSI and MSI fusion. It uses several advanced techniques, including network-in-network convolutional unit, batch normalization, parametric ReLU and skip connection. HR-HSI is gradually reconstructed by fusion of LR-HSI and MSI at different scales. In the meanwhile, we propose the RAP loss to simultaneously measure spatial and spectral distortions. The experiments demonstrate that our proposed network outperforms several state-of-the-art methods with regard to evaluation metrics and visual inspection.

REFERENCES

[1] J. Liu and J. Zhang, "Spectral unmixing via compressive sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 7099–7110, Nov. 2014.

[2] N. Audebert, B. Le Saux, and S. Lefevre, "Deep learning for classification of hyperspectral data: A comparative review," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 159–173, Jun. 2019.

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

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

[5] Y. Zhou, L. Feng, C. Hou, and S.-Y. Kung, "Hyperspectral and multispectral image fusion based on local low rank and coupled spectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5997–6009, Oct. 2017.

[6] C.-H. Lin, F. Ma, C.-Y. Chi, and C.-H. Hsieh, "A convex optimization-based coupled nonnegative matrix factorization algorithm for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1652–1667, Mar. 2018.

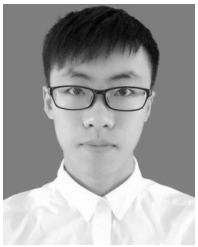
[7] F. Zhou, R. Hang, Q. Liu, and X. Yuan, "Pyramid fully convolutional network for hyperspectral and multispectral image fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 5, pp. 1549–1558, May 2019.

[8] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[10] C. Lanaras, E. Baltsavias, and K. Schindler, "Hyperspectral super-resolution by coupled spectral unmixing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 3586–3594.

[11] M. Eismann and R. Hardie, "Hyperspectral resolution enhancement using high-resolution multispectral imagery with arbitrary response functions," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 455–465, Mar. 2005.
[12] Q. Wei, N. Dobigeon, and J.-Y. Tourneret, "Fast fusion of multi-band images based on solving a Sylvester equation," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4109–4121, Nov. 2015.
[13] Q. Wei, N. Dobigeon, J.-Y. Tourneret, J. Bioucas-Dias, and S. Godsill, "R-FUSE: Robust fast fusion of multiband images based on solving a Sylvester equation," *IEEE Signal Process. Lett.*, vol. 23, no. 11, pp. 1632–1636, Nov. 2016.
[14] B. Lin, X. Tao, M. Xu, L. Dong, and J. Lu, "Bayesian hyperspectral and multispectral image fusions via double matrix factorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5666–5678, Oct. 2017.
[15] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.
[16] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1646–1654.
[17] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 5835–5843.
[18] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1874–1883.
[19] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, p. 594, Jul. 2016.
[20] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Multispectral and hyperspectral image fusion using a 3-D-convolutional neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 639–643, May 2017.
[21] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. Zhang, "A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 978–989, Mar. 2018.
[22] Z. Shao and J. Cai, "Remote sensing image fusion with deep convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1656–1669, May 2018.
[23] J. Yang, Y.-Q. Zhao, and J. Chan, "Hyperspectral and multispectral image fusion via deep two-branches convolutional neural network," *Remote Sens.*, vol. 10, no. 5, p. 800, May 2018.
[24] X. Liu, Y. Wang, and Q. Liu, "Remote sensing image fusion based on two-stream fusion network," in *Proc. 24th Int. Conf. MultiMedia Modeling (MMM)*, Bangkok, Thailand, Feb. 2018, pp. 428–439.
[25] R. Dian, S. Li, A. Guo, and L. Fang, "Deep hyperspectral image sharpening," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5345–5355, Nov. 2018.
[26] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. e3, 2016. [Online]. Available: http://distill.pub/2016/deconv-checkerboard
[27] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis.-(ECCV)*, Amsterdam, The Netherlands, Oct. 2016, pp. 391–407.
[28] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.-(ECCV)*, Amsterdam, The Netherlands, Oct. 2016, pp. 694–711.
[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015.
[30] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 105–114.
[31] V. J. Hill, R. C. Zimmerman, W. P. Bissett, H. Dierssen, and D. D. R. Kohler, "Evaluating light availability, seagrass biomass, and productivity using hyperspectral airborne remote sensing in Saint Joseph's Bay, Florida," *Estuaries Coasts*, vol. 37, no. 6, pp. 1467–1489, Nov. 2014.
[32] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogramm. Eng. Remote Sens.*, vol. 63, no. 6, pp. 691–699, 1997.
[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015.
[34] J. Yang, Y. Zhao, C. Yi, and J. C.-W. Chan, "No-reference hyperspectral image quality assessment via quality-sensitive features learning," *Remote Sens.*, vol. 9, no. 4, p. 305, Mar. 2017.



Shuang Xu is currently pursuing the Ph.D. degree in statistics with the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China.

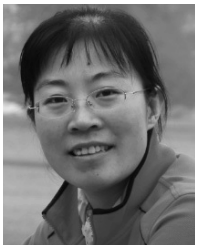
His research interests include Bayesian statistics, deep learning, and complex networks.

Ouafa Amira is currently pursuing the Ph.D. degree in statistics with the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China.

Her research interests include optimization, clustering, and deep learning.

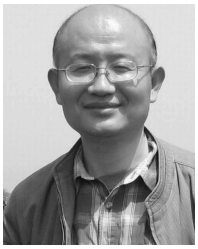
Junmin Liu (Member, IEEE) received the M.S. degree in computational mathematics from Ningxia University, Yinchuan, China, in 2009, and the Ph.D. degree in applied mathematics from Xi'an Jiaotong University, Xi'an, China, in 2013.

He is currently an Associate Professor with the School of Mathematics and Statistics, Xi'an Jiaotong University. His research interests include hyperspectral unmixing, remotely sensed image fusion, and deep learning.



Chun-Xia Zhang received the Ph.D. degree in applied mathematics from Xi'an Jiaotong University, Xi'an, China, in 2010.

She is currently an Associate Professor with the School of Mathematics and Statistics, Xi'an Jiaotong University. She has authored or coauthored about 30 journal articles in ensemble learning techniques and nonparametric regression. Her main interests are in the area of ensemble learning, variable selection, and deep learning.



Jiangshe Zhang was born in 1962. He received the M.S. and Ph.D. degrees in applied mathematics from Xi'an Jiaotong University, Xi'an, China, in 1987 and 1993, respectively.

He is currently a Professor with the Department of Statistics, Xi'an Jiaotong University. He has authored or coauthored one monograph and over 80 conference and journal publications in robust clustering, optimization, short-term load forecasting for electric power systems, and remote sensing image processing. His research interests include Bayesian statistics, global optimization, ensemble learning, and deep learning.

Guanghai Li was born in 1970. He received the master's and Ph.D. degrees in materials processing engineering from the South China University of Technology, Guangzhou, China, in 1998 and 2002, respectively.

Since 2005, he has been with the China Special Equipment Inspection and Research Institute, Beijing, China. He has authored or coauthored two monographs and over 50 conference and journal publications in nondestructive testing technology. His research interests include special equipment data analysis and fault diagnosis.