





A Twice Optimizing Net With Matrix Decomposition for Hyperspectral and Multispectral Image Fusion

Dunbin Shen , Jianjun Liu , *Member, IEEE*, Zhiyong Xiao , Jinlong Yang, and Liang Xiao , *Member, IEEE*

Abstract—Fusing a low-resolution hyperspectral (LRHS) image and a high-resolution multispectral (HRMS) image to generate a high-resolution hyperspectral (HRHS) image has grown a significant and attractive application in remote sensing fields. Recently, the popularization of deep learning has injected more possibilities into the fusion work. However, there still exists a difficulty that is how to make the best of the acquired LRHS and HRMS images. In this article, we present a twice optimizing net with matrix decomposition to fulfill the fusion task, which can be roughly divided into three stages: pre-optimization, deep prior learning, post-optimization. Specifically, we first transform this fusion problem into a spectral optimization problem and a spatial optimization problem with the help of matrix decomposition. These two optimization problems can be handled sequentially by solving a linear equation, respectively, and then we can obtain the initial HRHS image by multiplying the two solutions. Next, we establish the mapping between the initial image and the reference image through an end-to-end deep residual network based on local and nonlocal connectivity. In order to get better performance, we have customized a loss function specifically for the fusion task as well. Finally, we return the predicted result again to the optimization procedure to get the final fusion image. After the evaluation on three simulated datasets and one real dataset, it illustrates that the proposed method outperforms many state-of-the-art ones.

Index Terms—Convolutional neural network (CNN), hyperspectral image, image fusion, loss function, super resolution.

I. INTRODUCTION

HYPERSPECTRAL images (HSIs) with hundreds of bands, which contain sufficient spectral characteristics are widely used in many remote sensing fields [1]–[8], such as military surveillance, farming, geographic information monitoring and weather report. In order to do analysis more precisely and make decisions more appropriately, HSIs had better possess resolution as high as possible. While owing to the deficiencies of current satellite sensors, it seems impossible to acquire high-resolution hyperspectral (HRHS) images directly. Fortunately,

Manuscript received May 1, 2020; revised June 23, 2020; accepted July 10, 2020. Date of publication July 15, 2020; date of current version July 27, 2020. This work was supported by the National Natural Science Foundation of China under Grant 61601201. (Corresponding author: Jianjun Liu.)

Dunbin Shen, Jianjun Liu, Zhiyong Xiao, and Jinlong Yang are with the Jiangsu Provincial Engineering Laboratory for Pattern Recognition and Computational Intelligence, Jiangnan University, Wuxi 214122, China (e-mail: sdb_2012@163.com; liuofficial@163.com; zhiyong.xiao@jiangnan.edu.cn; yjgedeng@163.com).

Liang Xiao is with the School of Computer Science and Engineering, the Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: xiaoliang@mail.njust.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2020.3009250

low-resolution hyperspectral (LRHS) and high-resolution multispectral (HRMS) images can be conveniently obtained by nowadays imaging sensors, respectively. As a result, using HRMS and LRHS images to integrate desired HRHS images via a list of algorithms have become an appropriate and efficient option in research front. Recently, compared with traditional methods, deep learning has achieved overwhelming performance in many fields especially in image processing. Convolutional neural network (CNN), by virtue of its local connectivity and weight sharing properties, is widely used in HSI classification [9]–[13]. For instance, Chen *et al.* [10] applied CNN for HSI classification to extract the spectral-spatial features and enhanced the performance. Besides the classification task, CNN also performs outstandingly in single HSI super-resolution (SHSISR) [14]–[18], which aims to reconstruct a high-resolution image only by a single low-resolution image. For example, Li *et al.* [14] combined a spatial constraint strategy with a deep spectral difference CNN to acquire images with high spatial-resolution while protecting the spectral information. Yamanaka *et al.* [15] proposed a deep CNN with skip connection and network in network (DCSCN) to extract features and reconstruct details, which achieved outstanding performance in both accuracy and running time.

Unlike the SHSISR, the LRHS and HRMS image fusion contains two inputs, which indicates its difficult to fully integrate them. To deal with this problem, there has appeared two kinds of methods. One is to simply concatenate the upsampled LRHS image and the HRMS image into a whole and then feed it as input to the CNN to obtain the HRHS image. However, these methods will lead to the size inequality of input and output, which is not beneficial for the network design and can limit the speed of training. The other is using two separate networks to extract spectral and spatial features, respectively, and then fusing the two kinds of features to reconstruct the HRHS image. Not only are the spectral and spatial features hard to extract separately, but also the information distortion in the features fusion is difficult to control. As is known to all, image prior modeling can translate the fusion problem into an optimization problem constrained by HSI priors. In this article, we adopt a way based on image priors to acquire HRHS image via a twice optimizing net with matrix decomposition (TONWMD), as shown in Fig. 1. With the pre-optimization, we obtain a composite image, which has the same size as the desired HRHS image and contains spectral and spatial information, as rich as possible. Taking the composite image as the input and the reference image as the target, a deep residual network based on local and nonlocal connectivity is

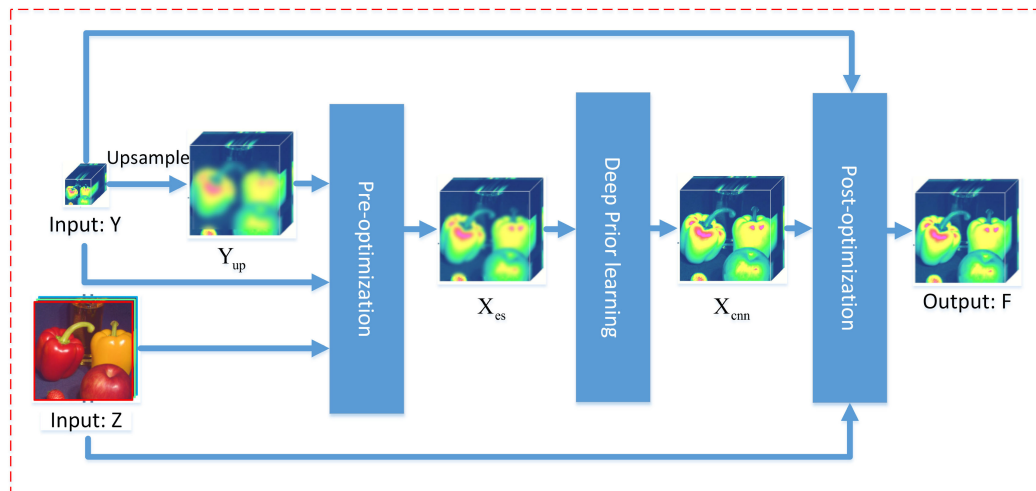


Fig. 1. Scheme of the proposed TONWMD method.

established to train the mapping model. When the model can achieve excellent performance, we feed the composite image from the test data to the model to predict the HRHS image. Since the predicted image can still be improved, we return it to the post-optimization to get the final fusion image. Our contributions are described briefly as follows.

- 1) We formulate a complicated fusion problem into a spectral optimization problem and a spatial optimization problem with the help of matrix decomposition. These two optimization problems can be simply handled sequentially by solving a linear equation, respectively. To get better performance, the optimization procedure is executed a total of twice before and after the prior learning. The pre-optimization is helpful for the design and training of the network while preserving both the spatial and spectral information well. The post-optimization can enhance the final performance.
- 2) To learn priors, we propose an effective deep residual network, which considers the long-distance dependency, the features of different levels and the multiscale analysis.
- 3) We also create a special loss function for this fusion work to achieve further performance, which covers three aspects of space, spectrum, and structure.
- 4) Experimental results evaluated on three simulated datasets and one real dataset have demonstrated our TONWMD method can achieve the state-of-the-art results both visually and quantitatively.

The remainder of this article is organized as follows. In Section II, we give a brief review of the hyperspectral and multispectral fusion methods. The proposed TONWMD method and its blind version are introduced in Section III. In Section IV, the experimental results and corresponding discussions on three simulated datasets and one real dataset are presented. Conclusion and future research directions are given in Section V.

II. RELATED WORK

HSI fusion algorithms can be roughly divided into three categories: pan-sharpening-based methods [19]–[22], matrix

factorization-based methods [23]–[30], and deep learning-based methods [31]–[39].

Pan-sharpening [40]–[43] is to obtain the HRHS image by fusing the LRHS image with a high-resolution panchromatic image, which mainly includes component substitution (CS), multiresolution analysis (MRA), Bayesian-based approaches, unmixing-based methods, and so on. Recently, some approaches like CS and MRA for pan-sharpening has been introduced to the LRHS and HRMS image fusion problem. Typically, Aiazzi *et al.* [19] considered the effect of spectral response function in original CS procedure, which could be extended to the fusion of LRHS and HRMS images via constructing pan-sharpening sub-problems. Each subproblem was to integrate one HRMS band with its corresponding LRHS bands. Selva *et al.* [21] build a linear regression model between each band of the LRHS image and all bands of the HRMS image to get the desired HRHS image. Due to the low spectral resolution of the panchromatic image, large spectral distortions caused by these methods are inevitable.

The essence of matrix factorization-based methods is to assume that the HRHS image contains a small number of pure spectral signatures and it can be estimated by multiplying the spectral basis with the corresponding coefficients. Kawakami *et al.* [23] utilized a sparse prior to learn the spectral basis from LRHS image and then conducted sparse coding on the HRMS image to get the coefficients. Furthermore, the priors of spectral unmixing were used to regularize the fusion problem. For instance, Yokoya *et al.* [24] adopted nonnegative matrix factorization-based spectral unmixing to learn endmembers and abundance from LRHS and HRMS images separately. In addition, to make full use of the nonlocal spatial similarities of the HRHS image, in [25], a nonnegative dictionary-learning algorithm was proposed to learn the spectral basis and the structured sparse coding method was utilized to fit the coefficients. However, since the inherent spectral-spatial correlations of HSI are hard to fully exploited, tensor-based approaches were proposed to handle this issue. Based on the Tucker decomposition [26], Dian *et al.* [27] came up with the nonlocal sparse tensor factorization and then Li *et al.* [28] put forward the coupled sparse tensor factorization to address the fusion task.

Recently, deep learning-based approaches have been widely exploited for the HRMS and LRHS image fusion. Up to now, these approaches can be roughly summarized into three categories: input-level fusion, feature-level fusion, and prior-exploited fusion. In the first kind, the LRHS and HRMS images are simply concatenated into a whole as the input of CNN to get the desired HRHS image. Typically, Masi *et al.* [31] directly stacked the LRHS and HRMS images as a whole in the spectral dimension and subsequently fed it into the super-resolution CNN framework. Furthermore, in [32], the spatially decimated HRMS image was concatenated with the spatially resampled principle components of LRHS image as an input fed into a 3-D CNN. In order to extract multiscale spatial features, Yuan *et al.* [33] proposed a multiscale and multidepth convolution block. In the second kind, the spectral features contained in the LRHS image as well as the spatial features contained in the HRMS image are extracted, respectively. Subsequently, these two kinds of features will be further integrated to reconstruct the HRHS image. For example, a two-branch CNN in [35] was designed to extract the spatial and spectral features separately. Besides, Yang *et al.* [36] utilized a 1-D CNN to learn spectral features from the LRHS image and a 2-D CNN to learn spatial features from the HRMS image, respectively. To gradually reconstruct the HRHS image in the spatial domain from a global level to local level, Zhou *et al.* [37] proposed a pyramid fully convolutional network where a latent image containing spectral information was obtained by the encoder subnetwork and then the HRMS image was gradually integrated with the latent image in the form of gaussian pyramid. Different from the first two kinds, the prior-exploited fusion takes the prior knowledge including traditional priors and learned priors into consideration to formulate the fusion problem into an optimization problem. The optimization problem will be dealt with by different algorithms such as solving equations or iterative algorithms. For instance, Dian *et al.* [38] presented a deep HSI sharpening method (DHSIS), which utilized a deep CNN to learn the priors and the learned priors were returned to the optimization framework. Xie *et al.* [39] created a novel multispectral and hyperspectral fusion net (MHF-net), which exploits the approximate low-rankness prior to reduce the spectral distortions. And the iterative algorithm unfolded in a deep CNN was designed to solve the fusion model.

III. PROPOSED METHOD

A. Problem Formulation

In this article, the LRHS image is denoted by $\mathbf{Y} \in \mathbb{R}^{S \times n}$, where S is the number of bands and n is the number of pixels in each band. Correspondingly, $\mathbf{Z} \in \mathbb{R}^{s \times N}$ is the HRMS image with s and N being its band number and pixel number ($s < S, n < N$), $\mathbf{X} \in \mathbb{R}^{S \times N}$ is the target HRHS image.

Based on the generation of \mathbf{Y} and \mathbf{Z} , the well-known observed models can be written as follows:

$$\mathbf{Y} = \mathbf{X}\mathbf{B}\mathbf{D} + \mathbf{N}_y \quad (1)$$

$$\mathbf{Z} = \mathbf{R}\mathbf{X} + \mathbf{N}_z \quad (2)$$

where $\mathbf{B} \in \mathbb{R}^{N \times N}$ is the cyclic convolution operator, $\mathbf{D} \in \mathbb{R}^{N \times n}$ is the down-sampling matrix, $\mathbf{R} \in \mathbb{R}^{s \times S}$ is the spectral response of sensor, and \mathbf{N}_y and \mathbf{N}_z denote the noises contained in \mathbf{Y} and \mathbf{Z} , respectively.

B. Pre-optimization

There is no doubt that the fusion goal is to obtain the HRHS image \mathbf{X} via \mathbf{Y} and \mathbf{Z} . However, even though deep CNN is effective in learning priors from the existed HRHS images, the two inputs \mathbf{Y} and \mathbf{Z} with different sizes are hard to map with \mathbf{X} directly. Therefore, we prepare to obtain an initial HRHS image \mathbf{X}_{es} at first via fully utilizing \mathbf{Y} and \mathbf{Z} and then it will be used as the input of the deep CNN. Based on the imaging models (1) and (2), the fusion problem can be translated into an optimization problem

$$\min_{\mathbf{X}_{es}} \|\mathbf{Z} - \mathbf{R}\mathbf{X}_{es}\|_F^2 + \|\mathbf{Y} - \mathbf{X}_{es}\mathbf{B}\mathbf{D}\|_F^2 + \lambda \|\mathbf{X}_{es} - \mathbf{Y}_{up}\|_F^2 \quad (3)$$

where $\mathbf{Y}_{up} \in \mathbb{R}^{S \times N}$ is the up-sampled version of \mathbf{Y} produced by bicubic interpolation, $\|\cdot\|_F$ denotes the Frobenius norm and λ is the regularization parameter. As in [38], this problem can be handled by solving the Sylvester equation.

While the calculation process of the Sylvester equation is a little complicated and it may take more time, the better way is to divide the optimization problem (3) into two steps: the spatial optimization and the spectral optimization. As the simple division can not achieve global optimal effects, we introduce the matrix factorization acted on \mathbf{X}_{es} to handle the issue. The two submatrices obtained by the factorization will be optimized, respectively, and then multiplied to estimate the initial HRHS image. Specifically, the decomposition of \mathbf{X}_{es} can be written as

$$\mathbf{X}_{es} = \mathbf{P}\mathbf{A} \quad (4)$$

where $\mathbf{P} \in \mathbb{R}^{S \times c}$ is the matrix composed of c orthogonal basis and $\mathbf{A} \in \mathbb{R}^{c \times N}$ is the transition matrix. Considering the spectral correlations of \mathbf{X}_{es} and \mathbf{Y}_{up} , we can get the initial \mathbf{P} from \mathbf{Y}_{up} via the singular value decomposition (SVD)

$$\mathbf{U}, \mathbf{\Sigma}, \mathbf{P}^T = \text{svds}(\mathbf{Y}_{up}^T, c) \quad (5)$$

where $\mathbf{U} \in \mathbb{R}^{N \times c}$ is the orthogonal matrix, $\mathbf{\Sigma} \in \mathbb{R}^{c \times c}$ is the diagonal matrix and svds is the SVD function, which saves the top c largest single values for dimension reduction without damaging the valid information. Based on models (1) and (2), we first optimize \mathbf{A} in the spectral field

$$\min_{\mathbf{A}} \|\mathbf{R}\mathbf{P}\mathbf{A} - \mathbf{Z}\|_F^2 + \lambda_1 \|\mathbf{P}\mathbf{A} - \mathbf{Y}_{up}\|_F^2 \quad (6)$$

and then optimize \mathbf{P} in the spatial field using the obtained \mathbf{A}

$$\min_{\mathbf{P}} \|\mathbf{P}\mathbf{A}\mathbf{B}\mathbf{D} - \mathbf{Y}\|_F^2 + \mu_1 \|\mathbf{P}\mathbf{A} - \mathbf{Y}_{up}\|_F^2 \quad (7)$$

where λ_1 and μ_1 are the positive parameters to balance the weight occupied by each term. The interpolated image \mathbf{Y}_{up} is utilized twice to ensure less distortion and better performance.

Since (6) and (7) are both quadratic convex optimization problem, they have unique solution. By forcing the derivation of (6) for \mathbf{A} and (7) for \mathbf{P} to be zero, respectively, we create two

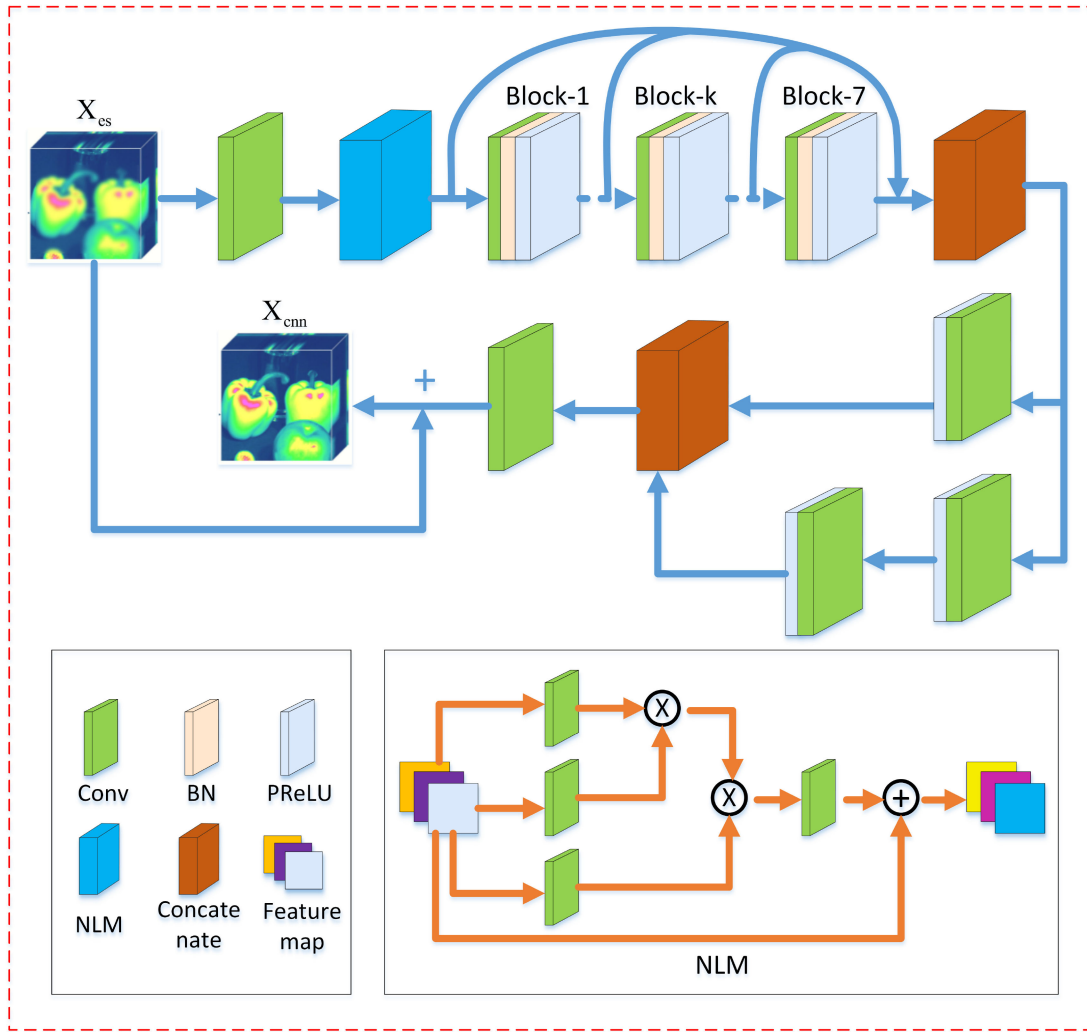


Fig. 2. Architecture of the proposed TONWMD network.

linear equations as follows:

$$\mathbf{H}_1 \mathbf{A} = \mathbf{H}_2 \quad (8)$$

$$\mathbf{P} \mathbf{H}_3 = \mathbf{H}_4 \quad (9)$$

where

$$\mathbf{H}_1 = (\mathbf{R}\mathbf{P})^T(\mathbf{R}\mathbf{P}) + \lambda_1 \mathbf{P}^T \mathbf{P}$$

$$\mathbf{H}_2 = (\mathbf{R}\mathbf{P})^T \mathbf{Z} + \lambda_1 \mathbf{P}^T \mathbf{Y}_{\text{up}}$$

$$\mathbf{H}_3 = (\mathbf{A}\mathbf{B}\mathbf{D})(\mathbf{A}\mathbf{B}\mathbf{D})^T + \mu_1 \mathbf{A}\mathbf{A}^T$$

$$\mathbf{H}_4 = \mathbf{Y}(\mathbf{A}\mathbf{B}\mathbf{D})^T + \mu_1 \mathbf{Y}_{\text{up}} \mathbf{A}^T. \quad (10)$$

To ensure (8) and (9) are not ill-posed, c should satisfy that $c < S$ and $c < N$. The solutions are listed as

$$\mathbf{A} = \mathbf{H}_1^{-1} \mathbf{H}_2 \quad (11)$$

$$\mathbf{P} = \mathbf{H}_4 \mathbf{H}_3^{-1} \quad (12)$$

After obtaining the optimized \mathbf{A} and \mathbf{P} , we take them into (4) and get the initial HRHS image \mathbf{X}_{es} .

C. Network Architecture

After the preoptimization, we have gotten a HRHS image \mathbf{X}_{es} , which has the same size with the desired image \mathbf{X} . With \mathbf{X}_{es} as the input and \mathbf{X} as the target, we can build an end to end deep residual network. The overall architecture is shown in Fig. 2.

We adopt a CNN architecture based on DCSCN, which mainly includes three parts: nonlocal module (NLM) [44], skip connection, and inception [45]. The descriptions of the framework are as follows and the detailed configuration of the convolution operation in each layer is shown in Table I.

The first convolution without using activation is applied to extract shallow features serving for the next operations. The NLM is used to take advantage of the influence of other locations around the point to capture long-distance dependence when extracting features at a certain point, and obtain better feature representation. The pure convolution calculation is a local operation and does not care the intrinsic correlation of each pixel in the overall image. In order to capture the effect of long-distance dependence, the convolution layer is often piled

up to deepen the network, which will increase the amount of calculation and make the learning become difficult. Therefore, we add the NLM before the deep feature extraction network to achieve nonlocal and local combination, which does a great favor to the feature extraction and the final fusion.

We utilize skip connection in the feature extraction network, which considers the hierarchical nature of features. Traditional convolutional networks often end up using high-level features, ignoring the role of low-level features. High-level features are generally abstract features such as semantics, while low-level features mainly include outlines, boundaries, and so on. Obviously, low-level features are indispensable in image fusion. So we merge features from the NLM to the last layer in the feature extraction network to reduce information loss and facilitate convergence. In each layer, we use batch normalization [46] to avoid overfitting and accelerate the training process after the convolution. Besides, we adopt the parametric rectified linear unit (PReLU) [47] as the activation function for avoiding the “dying ReLU” problem caused by ReLU [48].

Inception is a parallel CNN structure, which is designed for image reconstruction via connecting feature maps produced by filters of different size. Considering the actual effect at the same time reducing parameters, we only use two lines in the parallel network. One line is the 1×1 CNN just for reducing the dimension because the cascaded features obtained after the skip connection have a large dimension, which will increase the computational requirements too much. The other line also uses 1×1 CNN for dimensional reduction at first and then 3×3 CNN to generate satisfying feature maps. The two lines are cascaded to reconstruct the details. This not only spreads the width of the network but also increases its nonlinearity, which is of great significance for the improvement of the quality of the fusion result.

The 1×1 CNN following the inception is for dimensional tuning, which adjusts the number of channels to the ones of the target image.

In order to solve the problem of gradient explosion or disappearing, an end-to-end residual connection is also used so that what the entire network learns is the residual, i.e., the high-frequency details. It can speed up the training as well as ensure the low-frequency information is not lost, which is conducive to a stable fusion effect.

D. Network Loss

As we know, the standard L1-norm or L2-norm is often adopted as the loss function in CNN-based image fusion and it has achieved certain performance. However, either the standard L1-norm or L2-norm can only measure the spatial difference between the output and the target images, which is not exactly suitable for the HRMS and LRHS image fusion. Compared with natural images, the reconstruction effect of HSIs should not only be reflected in spatial similarity, but also ensure the similarity in spectrum and the consistency in structure. Therefore, we tailor a specific loss for the fusion task considering three aspects of space, spectrum, and structure. The loss function applied to train

the proposed network is defined as

$$\text{Loss}(\mathbf{X}_{\text{cnn}}, \mathbf{X}) = L_{\text{spat}} + \eta_1 * L_{\text{spec}} + \eta_2 * L_{\text{stru}} \quad (13)$$

where \mathbf{X}_{cnn} is the output image of reconstruction network, \mathbf{X} is the corresponding target image, L_{spat} is the spatial loss, L_{spec} is the spectral loss, L_{stru} is the structural loss, and η_1 and η_2 are used to balance the three terms.

Since the L2-norm loss easily produces blurry predictions in images reconstruction tasks [9], [37], we adopt the L1-norm loss as the main body to evaluate the similarity between output and target images in the pixelwise spatial domain, which is more conducive to the quality improvement. The function is defined as follow:

$$L_{\text{spat}}(\mathbf{X}_{\text{cnn}}, \mathbf{X}) = \frac{1}{SHW} \|\mathbf{X} - \mathbf{X}_{\text{cnn}}\|_1 \quad (14)$$

where H and W are the height and width of the HRHS image, and $\|\cdot\|_1$ denotes the L1-norm.

Spectral angle mapper (SAM) is an indispensable quality evaluation index in HRMS and LRHS image fusion, indicating the spectral quality. Thus, we measure the spectral loss by calculating the SAM. It is defined as follow:

$$L_{\text{spec}}(\mathbf{X}_{\text{cnn}}, \mathbf{X}) = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W \arccos \left(\frac{\mathbf{X}(i, j) \cdot \mathbf{X}_{\text{cnn}}(i, j)}{\|\mathbf{X}(i, j)\|_2 \|\mathbf{X}_{\text{cnn}}(i, j)\|_2 + g} \right) \quad (15)$$

where $\mathbf{X}_{\text{cnn}}(i, j)$ and $\mathbf{X}(i, j)$ denote the spectral vector at the spatial position (i, j) of \mathbf{X}_{cnn} and \mathbf{X} , $\|\cdot\|_2$ denotes the L2-norm and g is a small positive constant to avoid the denominator being 0.

Structural similarity index (SSIM) is also an important image quality evaluation index in the super-resolution task, which measures image similarity from three aspects: brightness, contrast, and structure. The higher SSIM not only guarantees less loss of image information, but also outperforms in visual effects. So we calculate the SSIM related L2-norm to measure the structural loss. It is defined as follow:

$$L_{\text{stru}}(\mathbf{X}_{\text{cnn}}, \mathbf{X}) = \left(1 - \frac{1}{S} \sum_{k=1}^S \text{SSIM}(\mathbf{X}(k), \mathbf{X}_{\text{cnn}}(k)) \right)^2 \quad (16)$$

where $\mathbf{X}_{\text{cnn}}(k)$ and $\mathbf{X}(k)$ denote k th band of \mathbf{X}_{cnn} and \mathbf{X} .

E. Post-optimization

In the testing process, we can acquire a fusion image \mathbf{X}_{cnn} by feeding the initialized testing HRHS image \mathbf{X}_{es} into the well-trained CNN. Even though \mathbf{X}_{cnn} is very close to the reference image \mathbf{X} , it can continue to be optimized for better performance.

Similar to preoptimization, we first obtain the initial \mathbf{P} from \mathbf{X}_{cnn} via SVD

$$\mathbf{U}, \mathbf{\Sigma}, \mathbf{P}^T = \text{svds}(\mathbf{X}_{\text{cnn}}^T, c). \quad (17)$$

Next, \mathbf{P} is utilized to obtain \mathbf{A} by solving the following optimization problem:

$$\min_{\mathbf{A}} \|\mathbf{RPA} - \mathbf{Z}\|_F^2 + \lambda_2 \|\mathbf{PA} - \mathbf{X}_{\text{cnn}}\|_F^2 \quad (18)$$

where $\lambda_2 > 0$ is the regularization parameter. Then, \mathbf{A} is used to obtain the new \mathbf{P} by solving the following optimization problem:

$$\min_{\mathbf{P}} \|\mathbf{PABD} - \mathbf{Y}\|_F^2 + \mu_2 \|\mathbf{PA} - \mathbf{X}_{\text{cnn}}\|_F^2 \quad (19)$$

where $\mu_2 > 0$ is the regularization parameter. Finally, the final estimated HRHS image \mathbf{F} can be obtained by calculating $\mathbf{F} = \mathbf{PA}$.

F. Blind Fusion

Based on whether prior knowledge other than HRMS and LRHS images is used or not, we can divide the fusion methods into nonblind and blind ones. Actually, our TONWMD method has assumed that the blur matrix \mathbf{B} and the spectral response matrix \mathbf{R} is known, which is unrealistic in real scenarios. In other words, our method is nonblind and not suitable for the real applications. So in order to fuse real data and do comparisons with those blind fusion algorithms, which only utilize HRMS and LRHS images, we need to blind our method.

HySure [49] is an outstanding blind fusion method, which can estimate \mathbf{B} and \mathbf{R} from \mathbf{Y} and \mathbf{Z} via convex optimization based on two quadratic data-fitting terms and total variation regularization. This estimation algorithm performs very well on simulated datasets and a little worse on real datasets for the effect of much noise. With the estimated \mathbf{B} and \mathbf{R} , our method can become a blind one (named BTONWMD).

IV. EXPERIMENTS

A. Data and Experimental Setup

In this article, experiments are conducted on three simulated datasets to evaluate the effectiveness of the proposed method: CAVE [50], Harvard [51], and University of Pavia (PU) [52]. We also use the University of Houston (UH) [53] as the real dataset to test the performance.

The CAVE dataset consists of 32 indoor HSIs, which was captured by the generalized assorted pixel camera with high quality. The HSIs have 31 bands where the wavelength range covers from 400 nm to 700 nm and each band has a spatial size of 512×512 . We use the first 20 images for training and the last 12 images for testing.

The Harvard dataset contains 50 images storing the indoor and outdoor scenes, which is about a number of objects, materials, and scale under daylight illumination. Each HSI has a spatial resolution of 1392×1040 and 31 spectral bands, of which the wavelength is ranging from 420–720 nm. We use the first 30 images for training and the last 20 images for testing.

The PU dataset is made up of 115 bands ranging from 430 to 860 nm. Each band of the HSI has a spatial size of 610×340 with a resolution of 1.3 m per pixel. There are 103 valid bands left in our experiment after removing the water vapor absorption and noise bands. As there is only one HSI, we crop a 128×128 subimage for testing and two nonoverlapping 128×128 subimages from the rest for training.

The UH dataset is released by the 2018 Data Fusion Contest of the IEEE Geoscience and Remote Sensing Society, which contains a HRMS image (RGB image) of size $83440 \times 24040 \times$

3 and a LRHS image of size $4172 \times 1202 \times 48$. For simplicity and consistency, we resize the HRMS image to size $33376 \times 9616 \times 3$. Thus, the ratio between the HRMS image and the LRHS image becomes 8. Like the PU data, we crop three nonoverlapping 1024×1024 subimages from the HRMS image and correspondingly three nonoverlapping 128×128 subimages from the LRHS image. We use the first two for training and the last one for testing. Since the ground truth is not available in the real dataset, we use Wald's protocol [54] to generate the training dataset. Specifically, we downsample the HRMS and LRHS images by 8 times as the training input. The original LRHS image is as the training label. While in testing, the original HRMS and LRHS images are as the input to predict the desired HRHS image.

For each of the three simulated datasets, LRHS images are acquired by applying a 8×8 Gaussian filter with a mean of 0 and a standard deviation of 2 and then downsampling it by 8 times. Taking different wavelength ranges and diversity into account, the way to produce the simulated HRMS images is different. For the CAVE and Harvard datasets, the spectral downsampling matrix \mathbf{R} comes from the response of a Nikon D700 camera, the HRMS images produced by which are RGB images. While for the PU dataset, we refer to the spectral response function of IKONOS, which describes the relationship between wavelength and multispectral channels. Specifically, we first select the points from the function, which has the same wavelength ranges with PU. As the number of selected points is less than the channels of PU, we then adopt the spline interpolation to spread the number of points to the number of spectral bands of PU. After removing the wavelength axis, the other axes of the points can form the spectral downsampling matrix \mathbf{R} . And the \mathbf{R} should be standardized at last. The HRMS image of PU is a multispectral image, which has five channels.

B. Implementation Details

When we estimate \mathbf{B} and \mathbf{R} via HySure, we set the number of nontruncated singular vectors as 10 to preserve necessary information while reducing dimensionality. And the effective size of the estimated \mathbf{B} are set as 10×10 . The other parameters are kept as the default values in the original algorithm.

In the preoptimization and postoptimization, the c is used in the SVD algorithm, which is related to the data compression and denoising. As the three simulated datasets have less noise, we just set c as the number of bands of HSIs to preserve the information. While the real dataset contains much noise, appropriate dimensionality reduction can denoise to achieve better performance. Thus, like in [49], we let $c = 10$ to preserve at least 99.95% of the energy of the original images in UH dataset.

In the training process, the input \mathbf{X}_{es} and the output \mathbf{X} of the CNN are cut into 32×32 patches. The stride size of cutting the CAVE and Harvard datasets is 16 while 1 for the PU and UH datasets. In addition, about 20% of the training slices are made up of the verification set to filter the model. The learning rate is initialized as 0.002 and every 10 epochs to decay a time by multiplying 0.5 until small than 10^{-7} . The weight decay is set to 0.0001 to suppress overfitting. For weight initialization, we use the method proposed in He *et al.* [48], which is a theoretically

suitable method for the network with PReLU. The coefficients η_1 and η_2 applied in the loss function are both set to 0.001 and the auxiliary parameter g is set to 0.001. The other parameters are set as the default values in Adam algorithm [55].

C. Compared Methods

We have compared the proposed method with five state-of-the-art approaches of HSI fusion: multiscale and multidepth convolutional neural network (MSDCNN) [33], remote sensing image fusion (RSIFNN) [35], 3-D CNN, DHSIS, and MHF-net. Among these methods, DHSIS is nonblind and the others are all blind.

D. Quality Measures

To evaluate the quality of fusion results, five popular indexes are used in this article.

- 1) Peak signal-to-noise ratio (PSNR): The PSNR is used to measure the average spatial similarities between the generated image and the reference image in all bands. The higher the value is, the less the spatial distortion is. The best value is ∞

$$\text{PSNR}(\mathbf{F}, \mathbf{X}) = \frac{1}{S} \sum_{k=1}^S \text{PSNR}(\mathbf{F}(\mathbf{k}), \mathbf{X}(\mathbf{k})). \quad (20)$$

- 2) SAM: The SAM indicates the spectral quality of the fusion image via calculating the angle averaged over the whole spatial domain. The smaller the degree is, the better the spectral quality is. The best value is 0

$$\begin{aligned} \text{SAM}(\mathbf{F}, \mathbf{X}) \\ = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W \arccos \left(\frac{\mathbf{F}(i, j) \cdot \mathbf{X}(i, j)}{\|\mathbf{F}(i, j)\|_2 \|\mathbf{X}(i, j)\|_2} \right). \end{aligned} \quad (21)$$

- 3) SSIM: The SSIM computes the average structural similarity in spatial domain between the generated image and the reference image. The higher the SSIM is, the more similar the spatial structure is. The ideal value is 1

$$\text{SSIM}(\mathbf{F}, \mathbf{X}) = \frac{1}{S} \sum_{k=1}^S \text{SSIM}(\mathbf{F}(\mathbf{k}), \mathbf{X}(\mathbf{k})). \quad (22)$$

- 4) Root-mean-squared error (RMSE): The RMSE is applied to represent the difference between the generated image and the reference image. Naturally, the smaller the difference is, the better the result is. The ideal value is 0

$$\begin{aligned} \text{RMSE}(\mathbf{F}, \mathbf{X}) \\ = \sqrt{\frac{1}{SHW} \sum_{k=1}^S \sum_{i=1}^H \sum_{j=1}^W (\mathbf{F}(i, j, k) - \mathbf{X}(i, j, k))^2}. \end{aligned} \quad (23)$$

- 5) Erreur relative globale adimensionnelle de synthse (ERGAS): The ERGAS is a global indicator that reflects overall fusion quality of the generated image, where d is the downsampling factor in the spatial domain, MSE is the mean square error function and MEAN is the mean value function. The best value of ERGAS is 0. And the

lower the value is, the better the overall quality is

$$\text{ERGAS}(\mathbf{F}, \mathbf{X}) = \frac{100}{d} \sqrt{\frac{1}{S} \sum_{k=1}^S \frac{\text{MSE}(\mathbf{F}(\mathbf{k}), \mathbf{X}(\mathbf{k}))}{\text{MEAN}^2(\mathbf{F}(\mathbf{k}))}}. \quad (24)$$

E. Parameters Selection

In our approach, the two parameters λ_1 and μ_1 are used in the preoptimization influencing the quality of \mathbf{X}_{cnn} while the other two parameters λ_2 and μ_2 are used in the postoptimization related to the quality of the final fusion result \mathbf{F} . Thus, the four parameters need to be set properly to obtain satisfying fusion effect. Considering the tuning convenience and consistency, we set $\lambda_1 = \mu_1 > 0$ and $\lambda_2 = \mu_2 > 0$.

It can be seen from Fig. 3(a) that as λ_1 increases, the PSNR value decreases on the three simulated datasets. The higher PSNR value of \mathbf{X}_{es} does not mean the better final result, which depends more on the latter prior learning. That is to say within reasonable limits, the value of λ_1 can be flexible. Therefore, we set $\lambda_1 = \mu_1 = 10^{-6}$ to ensure the overall quality on the three datasets. While in Fig. 3(b), the PSNR value first increases and later decreases on all three datasets with λ_2 increasing. As \mathbf{X}_{cnn} is very close to the ground truth \mathbf{X} , the fluctuations of PSNR values become very small. Taking the overall trend into consideration, we need an intermediate value. Thus, we set $\lambda_2 = \mu_2 = 0.002$ to obtain robust and satisfying fusion results.

F. Performance Comparison

Performance comparison with CAVE dataset. Table II shows the average objective results over 12 testing images in terms of PSNR, SAM, SSIM, RMSE, and ERGAS, where the optimal results are marked in blue among nonblind approaches and red among blind approaches for clarity. As is shown in this table, our proposed TONWMD method and its blind version can significantly outperform other competing methods with respect to all evaluation measures. It is suggested that our methods can better preserve both spatial and spectral information. In addition, in order to do comparison visually, the reconstructed images and the corresponding error images of the competing methods for the test image superballs (an HSI in the CAVE data) are displayed in Fig. 4. The reconstructed images come from the 31th band of the estimated images while the error images represent the differences between the reconstructed images and the ground truth. A meaningful region of each reconstructed image is marked and zoomed in 5 times for easy observation. It can be observed from the marked regions and the error images that our TONWMD and BTONWMD methods perform better in reconstructing the detailed structures and have less distortion than other methods. To further compare the fusion quality across different spectral bands, the PSNR curves of these methods are presented in Fig. 5(a). As can be seen from the picture, the TONWMD and BTONWMD methods still perform better in the most of spectral bands among the nonblind and blind methods, respectively.

Performance comparison with Harvard dataset. The average performance over 20 testing images of all competing methods

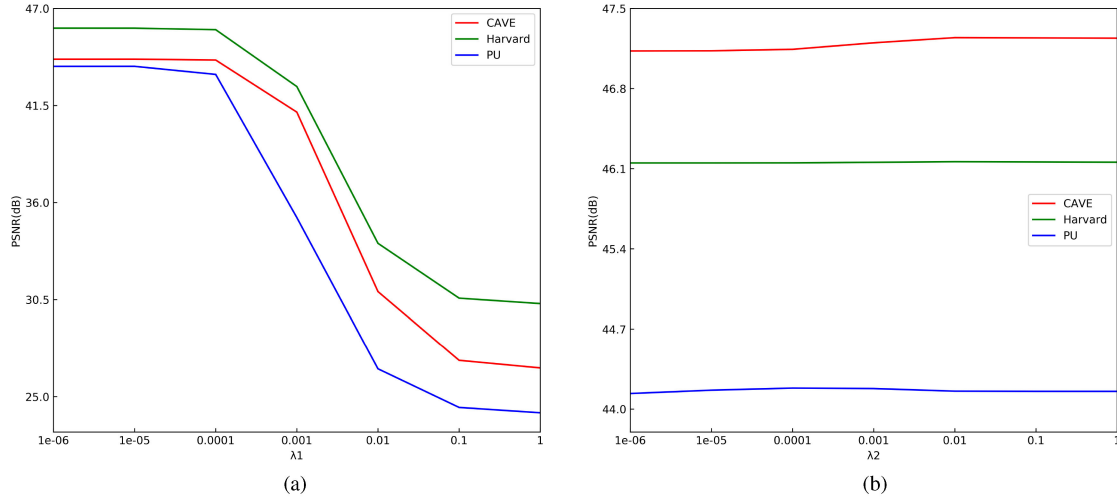


Fig. 3. (a) Average PSNR curves of \mathbf{X}_{es} as a function of λ_1 . (b) Average PSNR curves of \mathbf{F} as a function of λ_2 .

TABLE I
CONFIGURATION OF THE CONVOLUTION IN EACH LAYER

No.	Layer	Kernel Size	Batch-Norm	Activation
1	Conv[3×3]	3×3×S×64	False	-
2	Conv[1×1]	1×1×64×32	False	-
3	Conv[1×1]	1×1×64×32	False	-
4	Conv[1×1]	1×1×64×32	False	-
5	Conv[1×1]	1×1×32×64	False	-
6	Conv[3×3]	3×3×64×96	True	PReLU
7	Conv[3×3]	3×3×96×76	True	PReLU
8	Conv[3×3]	3×3×76×65	True	PReLU
9	Conv[3×3]	3×3×65×55	True	PReLU
10	Conv[3×3]	3×3×55×47	True	PReLU
11	Conv[3×3]	3×3×47×39	True	PReLU
12	Conv[3×3]	3×3×39×32	True	PReLU
13	Conv[1×1]	1×1×474×64	False	PReLU
14	Conv[1×1]	1×1×474×32	False	PReLU
15	Conv[3×3]	3×3×32×32	False	PReLU
16	Conv[1×1]	1×1×96×S	False	-

TABLE II
QUALITY MEASURES FOR CAVE DATASET

Type	Methods	PSNR	SAM	SSIM	RMSE	ERGAS
Blind	MSDCNN	40.29	5.69	0.9780	0.0103	1.2634
	RSIFNN	43.12	4.01	0.9879	0.0075	0.9134
	3D-CNN	43.64	4.15	0.9887	0.0071	0.8637
	MHF-net	46.05	3.92	0.9920	0.0053	0.6625
	BTONWMD	46.89	3.01	0.9938	0.0050	0.5891
Non-blind	DHSIS	46.26	3.06	0.9922	0.0053	0.6517
	TONWMD	47.25	2.85	0.9942	0.0048	0.5674

TABLE III
QUALITY MEASURES FOR HARVARD DATASET

Type	Methods	PSNR	SAM	SSIM	RMSE	ERGAS
Blind	MSDCNN	41.89	3.70	0.9737	0.0098	1.6748
	RSIFNN	43.82	3.33	0.9795	0.0081	1.3011
	3D-CNN	45.10	3.09	0.9836	0.0072	1.1380
	MHF-net	45.25	3.19	0.9834	0.0071	1.2370
	BTONWMD	46.06	2.91	0.9847	0.0066	1.0042
Non-blind	DHSIS	46.02	3.06	0.9838	0.0069	1.0088
	TONWMD	46.16	2.89	0.9849	0.0066	0.9886

on the Harvard dataset is reported in Table III. As the Harvard dataset is less challenging than the CAVE dataset, all the competing methods achieve good results, but our method and its blind version still perform better. Fig. 6 shows the reconstructed images and their corresponding error images of the competing methods for the cropped part of test image *img6* (an HSI in the Harvard data). A representative region of each reconstructed result is marked. It is obvious that the reconstructed images obtained by DHSIS, TONWMD, and BTONWMD methods are similar and they are very close to the ground truth. They achieve minimal reconstruction error at both the edges and smooth areas of the image. The PSNR curves as a function of the wavelengths of the spectral bands over the Harvard dataset for the test methods is shown in Fig. 5(b). It can be seen that the

DHSIS, TONWMD, and BTONWMD methods also perform better in most spectral bands than other competing methods.

Performance comparison with PU dataset. Table IV presents the competing performance of 1 testing image on the PU dataset. It can be found that our proposed TONWMD and its blind version outperform other competing methods in overall evaluation. Fig. 7 shows the reconstructed images and their corresponding error images of the competing methods for the test image. It can be observed that all competing methods have achieved satisfying results while the reconstruct details obtained by our TONWMD and BTONWMD methods are better. Fig. 5(c) displays the PSNR curves as a function of the wavelengths of the spectral bands over the PU dataset for the competing methods. As can

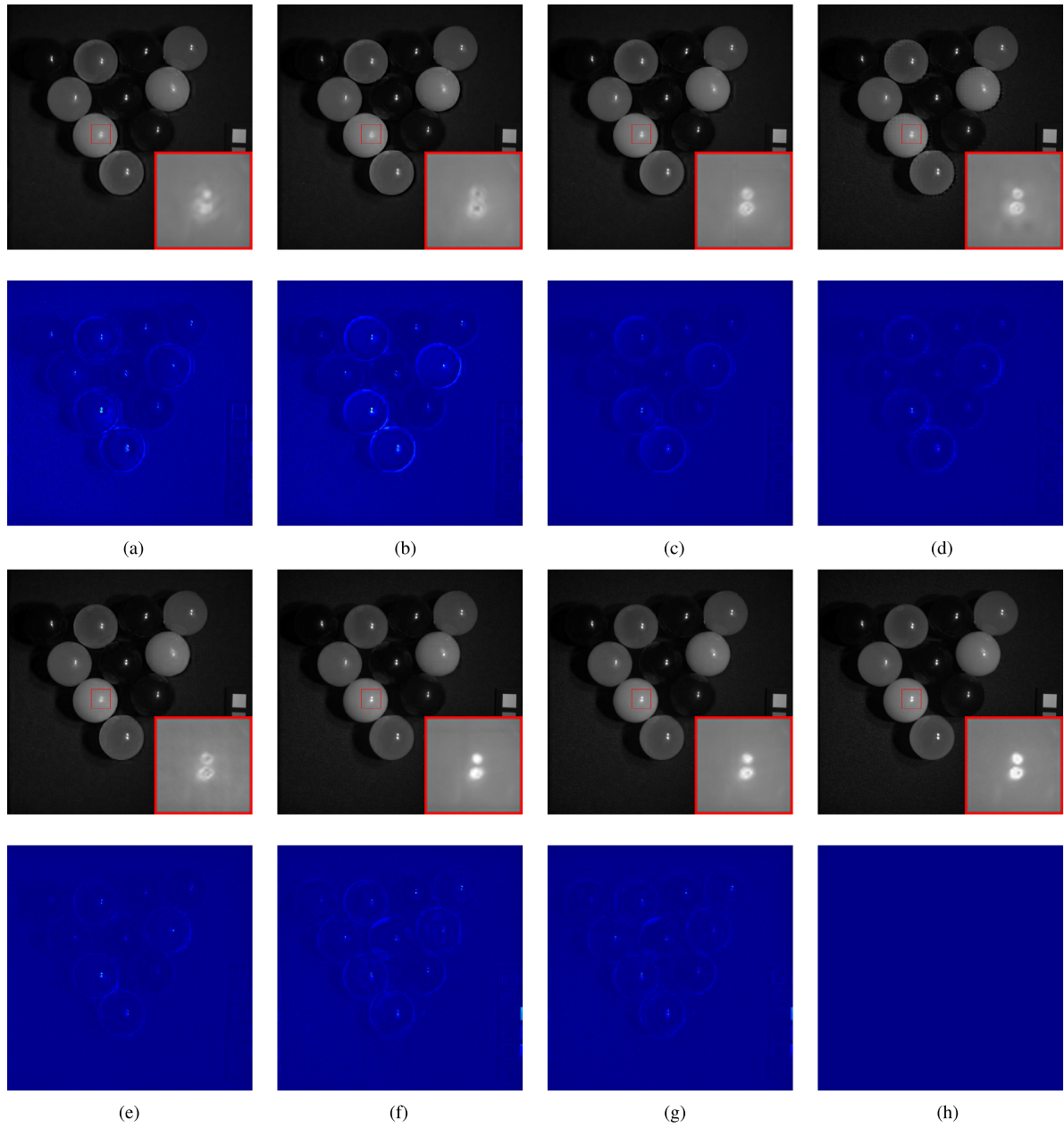


Fig. 4. Qualitative results of CAVE dataset at band 31. Top row: reconstructed images. Bottom row: reconstruction errors—light color indicates less error, dark color indicates larger error. (a) MSDCNN. (b) RSIFNN. (c) 3D-CNN. (d) DHSIS. (e) MHF-net. (f) BTONWMD. (g) TONWMD. (h) Ground truth.

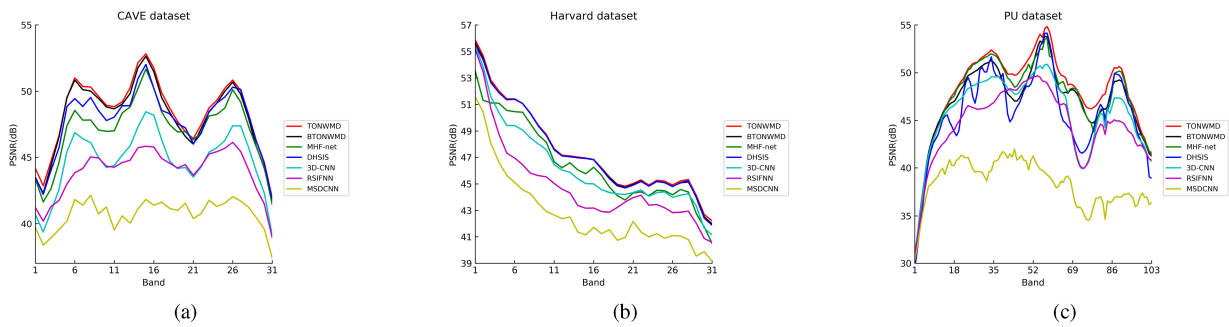


Fig. 5. Average PSNR curves as functions of the spectral bands for the test method. (a) CAVE dataset. (b) Harvard dataset. (c) PU dataset.

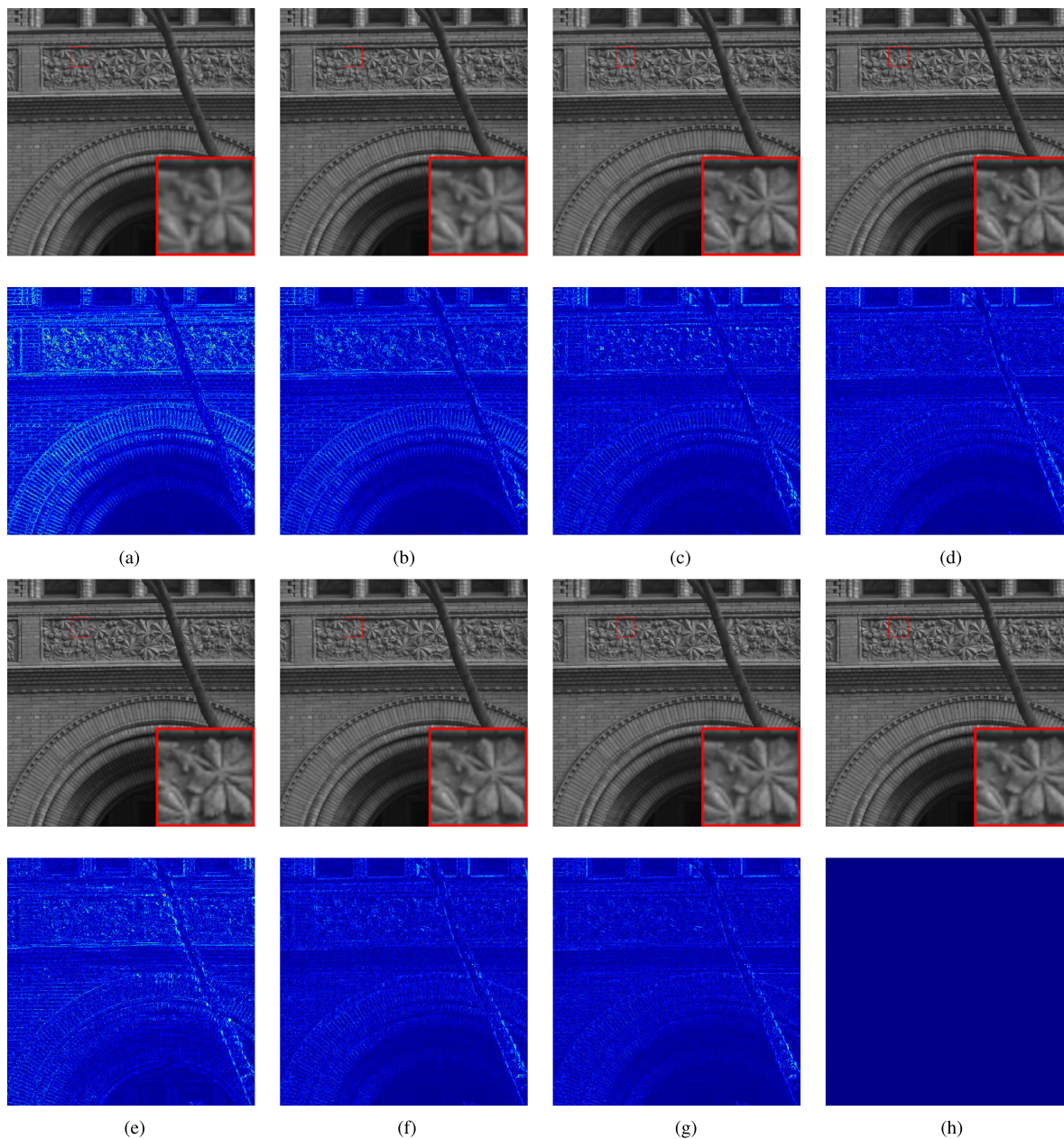


Fig. 6. Qualitative results of Harvard dataset at band 31. Top row: reconstructed images. Bottom row: Reconstruction errors—light color indicates less error, dark color indicates larger error. (a) MSDCNN. (b) RSIFNN. (c) 3D-CNN. (d) DHSIS. (e) MHF-net. (f) BTOWMD. (g) TONWMD. (h) Ground truth.

be seen from the picture, along with MHF-net method, our TONWMD and its blind version show higher PSNR value in most of the spectral bands than other competing methods from the overall perspective.

To further comparing the fusion quality, we can evaluate the classification performance over all the fusion results obtained by the competing methods on the PU dataset. As our test image does not contains enough classes, we select a new part of the original HSI to do classification. This selected part neither overlaps our training set nor our test set. The false color image and the ground reference map of the selected part are shown in Fig. 8. And there are 5 ground reference classes of interests shown in Table V. In our classification experiments, we build training sets by

randomly choosing 100 training pixels per class from the original HSI but not containing our selected part and use all the fusion results as the test sets. For simplicity, we adopt the classic support vector machines algorithm [56]. The classification performance can be seen from Table VI that our TONWMD method and its blind version as well as the 3D-CNN achieve higher accuracy than other competing methods. The reason why the classification accuracy is even higher than the original HRHS image is that the reconstruction work has denoised some noises.

Performance comparison with UH dataset. As nonblind methods require additional degradation information, which is unknown in this case, thus, we only compare our BTOWMD method with other blind methods. Fig. 9 shows a portion of the

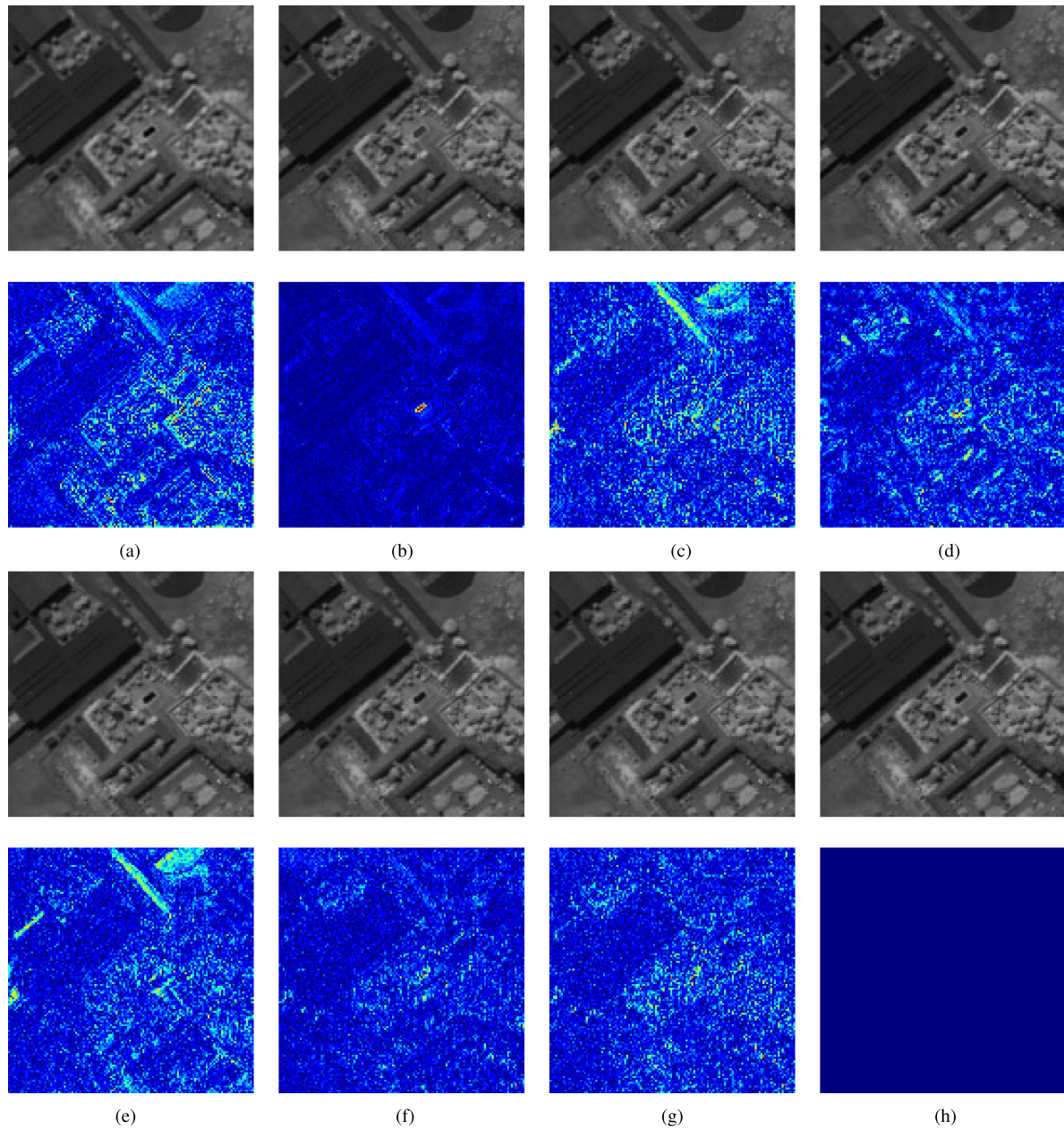


Fig. 7. Qualitative results of PU dataset at band 103. Top row: reconstructed images. Bottom row: Reconstruction errors—light color indicates less error, dark color indicates larger error. (a) MSDCNN. (b) RSIFNN. (c) 3D-CNN. (d) DHSIS. (e) MHF-net. (f) BTONWMD. (g) TONWMD. (h) Ground truth.

fusion results of the test data, which are RGB images generated by the evaluated spectral response function. Visual inspection evidently shows that the result obtained by our BTONWMD method is much closer to the ground truth with clearer details.

G. Effectiveness of Three Steps

In this article, our approach is made up of three steps, i.e., the preoptimization to obtain an initial HRHS image, using an innovative deep CNN to learn the mapping, and the postoptimization to further improve the performance. In order to illustrate the effectiveness of the three steps from objective perspective, average quantitative results of \mathbf{Y}_{up} , \mathbf{X}_{es} , \mathbf{X}_{cnn} , and

\mathbf{F} on the three data are displayed in Tables VII, VIII, and IX, respectively. It can be seen from the three tables that \mathbf{X}_{es} makes much better performance than \mathbf{Y}_{up} in all three datasets, which can prove the preoptimization is effective in preserving both spectral and spatial information while integrating two images into a whole initially. The result \mathbf{X}_{cnn} predicted by CNN also has better quantitative results on all three data compared with \mathbf{X}_{es} . It illustrates the prior learning is effective and helpful for the fusion work. In addition, the final fusion result \mathbf{F} achieves further raise based on \mathbf{X}_{cnn} over all three datasets, which means the postoptimization has further impact on the quality improving of the reconstruction task. In general, all the three steps are indispensable in getting better performance in this article.

TABLE IV
QUALITY MEASURES FOR PU DATASET

Type	Methods	PSNR	SAM	SSIM	RMSE	ERGAS
Blind	MSDCNN	37.71	3.60	0.9621	0.0130	1.2141
	RSIFNN	42.32	2.33	0.9808	0.0077	0.8284
	3D-CNN	43.17	2.24	0.9831	0.0069	0.7188
	MHF-net	43.70	2.18	0.9835	0.0065	0.7329
	BTONWMD	43.92	2.10	0.9839	0.0064	0.7096
Non-blind	DHSIS	42.46	2.49	0.9788	0.0075	0.8396
	TONWMD	44.15	2.07	0.9838	0.0062	0.7080

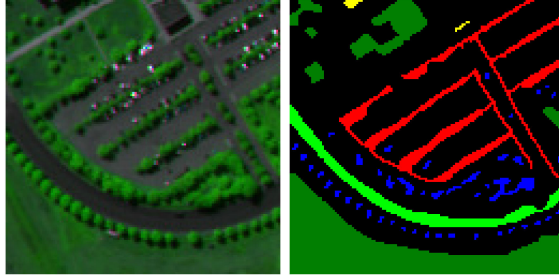


Fig. 8. Selected part of the PU dataset for classification. (a) RGB composite image of three bands (35,98,60). (b) Ground reference map.

TABLE V
FIVE GROUND REFERENCE CLASSES IN THE TEST IMAGE OF PU DATASET

NO	Name	Samples
C1	Asphalt	676
C2	Meadow	2927
C3	Trees	395
C4	Bricks	1232
C5	Shadows	33

H. Sensitivity to Noise

In order to demonstrate the robustness of our method to noise and further verify its effectiveness when facing noise environment, we add Gaussian white noise with different levels to the LRHS images and the HRMS images. To distinguish the spatial resolution of the LRHS images and the HRMS images, we intentionally set the decibel difference of the noise added to these two images as 10. Fig. 10 presents the comparison results on the CAVE dataset. We can find that the least sensitive to noise is the MSDCNN method although it has a relatively weak performance. Along with MHF-net, our TONWMD and BTONWMD methods are mildly influenced by decreasing the SNR of noise and they are steadily leading under different noise levels. That is to say, our TONWMD methods and its blind version have some robustness to noise and can be applied in practice.

TABLE VI
CLASSIFICATION PERFORMANCE ON PU DATASET IN OVERALL ACCURACY (OA), COHEN'S KAPPA (KA), AND AVERAGE PRODUCER'S ACCURACY (AA)

Methods	OA(%)	KA(%)	AA(%)
Bicubic	8.09	0.28	18.90
MSDCNN	55.77	44.36	70.30
RSIFNN	89.42	83.46	90.11
3-D CNN	90.91	85.59	90.05
DHSIS	90.77	85.33	89.72
MHF-net	89.11	82.95	88.90
BTONWMD	92.38	87.72	90.11
TONWMD	90.86	85.55	90.34
Ground truth	89.47	83.42	89.00

TABLE VII
AVERAGE QUANTITATIVE RESULTS OF \mathbf{Y}_{up} , \mathbf{X}_{es} , \mathbf{X}_{cnn} , AND \mathbf{F} ON CAVE DATASET

Methods	PSNR	SAM	SSIM	RMSE	ERGAS
\mathbf{Y}_{up}	26.34	7.27	0.7980	0.0520	6.2000
\mathbf{X}_{es}	44.12	3.94	0.9890	0.0067	0.7960
\mathbf{X}_{cnn}	46.07	2.97	0.9937	0.0055	0.6504
\mathbf{F}	47.25	2.85	0.9942	0.0048	0.5674

TABLE VIII
AVERAGE QUANTITATIVE RESULTS OF \mathbf{Y}_{up} , \mathbf{X}_{es} , \mathbf{X}_{cnn} , AND \mathbf{F} ON HARVARD DATASET

Methods	PSNR	SAM	SSIM	RMSE	ERGAS
\mathbf{Y}_{up}	29.89	4.11	0.8181	0.0344	4.2470
\mathbf{X}_{es}	45.88	2.96	0.9847	0.0067	1.0216
\mathbf{X}_{cnn}	46.09	2.89	0.9848	0.0066	0.9904
\mathbf{F}	46.16	2.89	0.9849	0.0066	0.9886

TABLE IX
AVERAGE QUANTITATIVE RESULTS OF \mathbf{Y}_{up} , \mathbf{X}_{es} , \mathbf{X}_{cnn} , AND \mathbf{F} ON PU DATASET

Methods	PSNR	SAM	SSIM	RMSE	ERGAS
\mathbf{Y}_{up}	24.04	8.82	0.4997	0.0628	5.0827
\mathbf{X}_{es}	43.72	2.15	0.9838	0.0065	0.6905
\mathbf{X}_{cnn}	43.95	2.12	0.9837	0.0063	0.7181
\mathbf{F}	44.15	2.07	0.9838	0.0062	0.7080

I. Computational Efficiency Comparison

To clarify the computational efficiency of our proposed method, we discuss the running time of each competing method on the four datasets. For the training phase, all the competing methods are implemented with TensorFlow [57] and run on a single GeForce GTX 1660 Ti 6 GB graphic card. Under the same circumstance, the MSDCNN takes about 4 h for training, the RSIFNN takes about 10 h, the 3-D CNN takes about 5 h, the DHSIS takes about 8 h, the MHF-net takes about 7 h while our TONWMD and BTONWMD methods take about 2 h. Based on the trained parameters, the fusion procedure is performed by using an Intel(R) Core(TM) i5-9300H CPU 2.40 GHz and a 8 GB RAM through Python 3.6. The average testing time of the competing methods are shown in Table X. By analyzing



Fig. 9. Qualitative results of UH dataset. (a) Bicubic. (b) MSDCNN. (c) RSIFNN. (d) 3D-CNN. (e) MHF-net. (f) BTONWMD. (g) Ground truth.

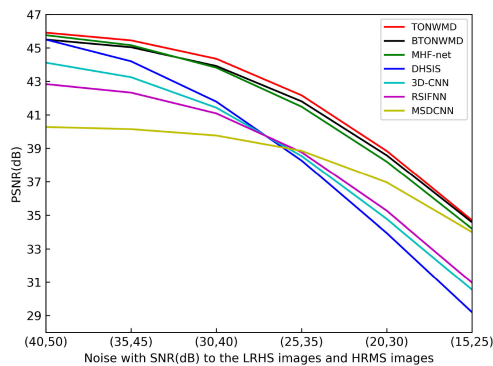


Fig. 10. Comparison of competing methods under different noise levels on CAVE dataset—smaller slope indicates better robustness.

TABLE X
RUNNING TIME (SECOND) OF THE FUSION METHODS

Methods	CAVE	Harvard	PU	UH	Average
MSDCNN	1.641	2.702	6.422	16.506	6.818
RSIFNN	1.395	1.920	3.575	8.783	3.918
3D-CNN	1.274	4.986	2.953	9.753	4.742
DHSIS	6.054	23.338	8.826	-	12.739
MHF-net	3.100	7.705	15.643	27.115	13.391
TONWMD	2.624	13.581	5.090	-	7.098
BTONWMD	2.824	17.423	5.240	14.204	9.923

the data in Table X, we can find that our TONWMD method is a little quicker than the DHSIS method. And the running speed of our BTONWMD method is slower than the RSIFNN, 3D-CNN, and MSDCNN methods while only quicker than the MHF-net method, which indicates the twice optimization and

TABLE XI

FURTHER PERFORMANCE IMPROVING BY USING THE TOOL ON CAVE DATASET

Methods	PSNR	SAM	SSIM	RMSE	ERGAS
MSDCNN	44.12	4.20	0.9885	0.0066	0.8113
RSIFNN	45.62	3.51	0.9921	0.0056	0.6807
3D-CNN	45.21	3.60	0.9915	0.0059	0.7150
DHSIS	46.71	2.93	0.9929	0.0050	0.6179
MHF-net	47.03	3.32	0.9932	0.0047	0.5899

the estimation of \mathbf{R} and \mathbf{B} have added some calculation. In fact, this little gap in running speed is within an acceptable range because the quality of the fusion has been improved.

J. Extension: General Tool for Performance Improving

Actually, the postoptimization part can be separated from our method into a general tool to help other methods improve performance. In this experiment, we let the fusion results of the competing methods directly pass our postoptimization without caring whether they are blind or unblind. To meet the actual situation, we use the estimated \mathbf{B} and \mathbf{R} . Table XI shows the experimental results of CAVE dataset. Compared with Table II, we can find that after using our postoptimization, the performance of the five competing methods has improved a lot. Besides, we also verify it on UH dataset. Combining Figs. 9 and 11, we can see that the image details become clearer and closer to the original RGB image after passing our postprocessing. It demonstrates that our postoptimization is a general tool and may serve for other HRMS and LRMS images fusion methods.

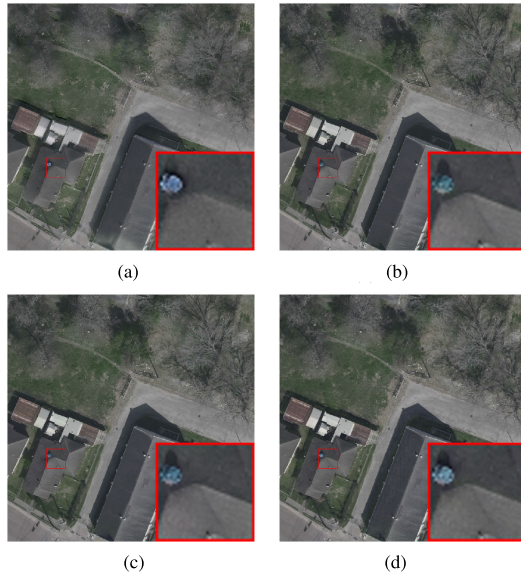


Fig. 11. Further performance improving by using the tool on UH dataset. (a) MSDCNN. (b) RSIFNN. (c) 3D-CNN. (d) MHF-net.

V. CONCLUSION

In this article, we propose a twice optimizing net based on imaging priors to tackle the HRMS and LRHS image fusion work. Different from other CNN-based methods, we first formulate the fusion problem into a spectral optimization problem and a spatial optimization problem by using matrix decomposition. After solving the two optimization problems sequentially, we obtain the initial HRHS image preserving both the spectral and spatial information well. Then, we design a deep residual network considering the features in different regions, levels and scales to learn the image priors, which can fully utilize the high nonlinearity to model the complex nonlinear relationship between the initial image and the target image. Besides, a special loss function is designed to help improve the performance. Finally, the predicted result obtained via CNN are sent to the post-optimization to further improve the performance. To fuse real datasets, we have gotten our blind version method with the help of HySure by estimating \mathbf{R} and \mathbf{B} from \mathbf{Y} and \mathbf{Z} before doing optimization. Experimental results on three simulated dataset and one real dataset demonstrate that our approach outperforms the state-of-the-art methods in both qualitative and quantitative ways.

In the future work, we will consider putting the twice optimization into the deep CNN to reduce the extra estimation work. Besides, the current network can be enhanced more effectively to further improve the performance.

ACKNOWLEDGMENT

The authors would like to thank Prof. A. Chakrabarti from Washington University for providing the Harvard dataset and Prof. P. Gamba from the University of Pavia for providing the PU dataset, they would also like to thank the National Center for Airborne Laser Mapping, the Hyperspectral Image Analysis

Laboratory, the University of Houston for providing the UH dataset, and also would like to thank the anonymous reviewers for their constructive comments on this article.

REFERENCES

- [1] N. Yokoya, C. Grohnfeldt, and J. Chanussot, "Hyperspectral and multi-spectral data fusion: A comparative review of the recent literature," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 2, pp. 29–56, Jun. 2017.
- [2] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.
- [3] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.
- [4] J. Liu, Z. Wu, L. Xiao, and H. Yan, "Learning multiple parameters for kernel collaborative representation classification," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–11, Jan. 2020, to be published, doi: [10.1109/TNNLS.2019.2962878](https://doi.org/10.1109/TNNLS.2019.2962878).
- [5] J. Liu, Z. Wu, L. Xiao, J. Sun, and H. Yan, "Generalized tensor regression for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 1244–1258, Feb. 2020.
- [6] J. Liu, Z. Wu, J. Li, A. Plaza, and Y. Yuan, "Probabilistic-kernel collaborative representation for spatial-spectral hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2371–2384, Apr. 2016.
- [7] R. A. Borsoi, T. Imbiriba, and J. C. M. Bermudez, "Super-resolution for hyperspectral and multispectral image fusion accounting for seasonal spectral variability," *IEEE Trans. Image Process.*, vol. 29, no. 1, pp. 116–127, Jul. 2019.
- [8] X. Han, J. Yu, J. Xue, and W. Sun, "Hyperspectral and multispectral image fusion using optimized twin dictionaries," *IEEE Trans. Image Process.*, vol. 29, no. 1, pp. 4709–4720, Mar. 2020.
- [9] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," in *Proc. Int. Conf. Learn. Representations*, Nov. 2016, pp. 1–14.
- [10] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.
- [11] M. E. Paoletti, J. M. Haut, R. Fernandezbeltran, J. Plaza, A. Plaza, and F. Pla, "Deep pyramidal residual networks for spectral-spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 740–754, Feb. 2019.
- [12] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "Hybrids: Exploring 3D–2D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2020.
- [13] J. Feng *et al.*, "CNN-based multilayer spatial-spectral feature fusion and sample augmentation with local and nonlocal constraints for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1299–1313, Apr. 2019.
- [14] Y. Li, J. Hu, X. Zhao, W. Xie, and J. Li, "Hyperspectral image super-resolution using deep convolutional neural network," *Neurocomputing*, vol. 266, pp. 29–41, 2017.
- [15] J. Yamanaka, S. Kuwashima, and T. Kurita, "Fast and accurate image super resolution by deep CNN with skip connection and network in network," in *Proc. Int. Conf. Neural Inf. Process.*, 2017, pp. 217–225.
- [16] J. M. Haut, R. Fernandezbeltran, M. E. Paoletti, J. Plaza, A. Plaza, and F. Pla, "A new deep generative network for unsupervised remote sensing single-image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6792–6810, Nov. 2018.
- [17] W. Xie, X. Jia, Y. Li, and J. Lei, "Hyperspectral image super-resolution using deep feature matrix factorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 6055–6067, Aug. 2019.
- [18] P. V. Arun, K. M. Buddhiraju, A. Porwal, and J. Chanussot, "CNN-based super-resolution of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. PP, pp. 1–16, Feb. 2020, to be published, doi: [10.1109/TGRS.2020.2973370](https://doi.org/10.1109/TGRS.2020.2973370).
- [19] B. Aiuzzi, S. Baronti, and M. Selva, "Improving component substitution pansharpening through multivariate regression of ms + pan data," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3230–3239, Oct. 2007.

- [20] Z. Chen, H. Pu, B. Wang, and G.-M. Jiang, "Fusion of hyperspectral and multispectral images: A novel framework based on generalization of pan-sharpening methods," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 8, pp. 1418–1422, Aug. 2014.
- [21] M. Selva, B. Aiazzi, F. Butera, L. Chiarantini, and S. Baronti, "Hyper-sharpening: A first approach on sigma data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 3008–3024, Jun. 2015.
- [22] X. Lu, J. Zhang, X. Yu, W. Tang, T. Li, and Y. Zhang, "Hyper-sharpening based on spectral modulation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 5, pp. 1534–1548, May 2019.
- [23] R. Kawakami, Y. Matsushita, J. Wright, M. Ben-Ezra, Y.-W. Tai, and K. Ikeuchi, "High-resolution hyperspectral imaging via matrix factorization," in *Proc. IEEE CVPR*, 2011, pp. 2329–2336.
- [24] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled non-negative matrix factorization (CNMF) for hyperspectral and multispectral data fusion: Application to pasture classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2011, pp. 1779–1782.
- [25] W. Dong *et al.*, "Hyperspectral image super-resolution via non-negative structured sparse representation," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2337–2352, May 2016.
- [26] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.
- [27] R. Dian, L. Fang, and S. Li, "Hyperspectral image super-resolution via non-local sparse tensor factorization," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 5344–5353.
- [28] S. Li, R. Dian, L. Fang, and J. M. Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4118–4130, Aug. 2018.
- [29] R. Dian and S. Li, "Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 5135–5146, Oct. 2019.
- [30] L. Sui, L. Li, J. Li, N. Chen, and Y. Jiao, "Fusion of hyperspectral and multispectral images based on a Bayesian nonparametric approach," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1205–1218, Apr. 2019.
- [31] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, pp. 594–615, Jul. 2016.
- [32] 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.
- [33] 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.
- [34] Z. He and L. Liu, "Hyperspectral image super-resolution inspired by deep laplacian pyramid network," *Remote Sens.*, vol. 10, no. 12, 2018, Art. no. 1939.
- [35] 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.
- [36] J. Yang, Y.-Q. Zhao, and J. C.-W. Chan, "Hyperspectral and multispectral image fusion via deep two-branches convolutional neural network," *Remote Sens.*, vol. 10, no. 5, pp. 800–822, 2018.
- [37] 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.
- [38] 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.
- [39] Q. Xie, M. Zhou, Q. Zhao, D. Meng, W. Zuo, and Z. Xu, "Multispectral and hyperspectral image fusion by MS/HS fusion net," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 1585–1594.
- [40] P. Liu, L. Xiao, J. Zhang, and B. Naz, "Spatial-hessian-feature-guided variational model for pan-sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2235–2253, Apr. 2016.
- [41] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "PanNet: A deep network architecture for pan-sharpening," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 1753–1761.
- [42] P. Liu, L. Xiao, and T. Li, "A variational pan-sharpening method based on spatial fractional-order geometry and spectral-spatial low-rank priors," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1788–1802, Mar. 2018.
- [43] Y. Zhang, C. Liu, M. Sun, and Y. Ou, "Pan-sharpening using an efficient bidirectional pyramid network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5549–5563, Aug. 2019.
- [44] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7794–7803.
- [45] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 1–9.
- [46] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, Mar. 2015, pp. 448–456.
- [47] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines vinod nair," in *Proc. Int. Conf. Mach. Learn.*, vol. 27, Jun. 2010, pp. 807–814.
- [48] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 1026–1034.
- [49] M. Simoes, J. Bioucas-Dias, L. B. Almeida, and J. Chanussot, "A convex formulation for hyperspectral image superresolution via subspace-based regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3373–3388, Jun. 2015.
- [50] F. Yasuma, T. Mitsunaga, D. Iso, and S. Nayar, "Generalized assorted pixel camera: Post-capture control of resolution, dynamic range and spectrum," *IEEE Trans. Image Process. : IEEE Signal Process. Soc.*, vol. 19, no. 6, pp. 2241–53, Mar. 2010.
- [51] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2011, pp. 193–200.
- [52] X. Huang and L. Zhang, "A comparative study of spatial approaches for urban mapping using hyperspectral rosis images over pavia city, northern italy," *Int. J. Remote Sens.*, vol. 30, no. 12, pp. 3205–3221, 2009.
- [53] Y. Xu *et al.*, "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1709–1724, Jun. 2019.
- [54] R. Carla, L. Santurri, B. Aiazzi, and S. Baronti, "Full-scale assessment of pansharpening through polynomial fitting of multiscale measurements," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6344–6355, Dec. 2015.
- [55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, Dec. 2015, pp. 1–15.
- [56] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, 1999.
- [57] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*.



Dunbin Shen received the B.S. degree in digital media technology from Hubei Minzu University, Enshi City, China, in 2016. He is currently working toward the M.D. degree in computer science and technology with the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China.

His research interests include deep learning and hyperspectral image fusion.



Jianjun Liu (Member, IEEE) received the B.S. degree in applied mathematics and the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, Nanjing, China, in 2009 and 2014, respectively.

He is currently an Associate Professor with the School of Artificial Intelligence and Computer Science, Jiangnan University, China. Since 2018, he has been a Postdoctoral Researcher with the Department of Electrical Engineering, City University of Hong Kong, China. His research interests include the areas of hyperspectral image classification, super-resolution, spectral unmixing, sparse representation, computer vision, and pattern recognition.



Zhiyong Xiao received the Ph.D. degree in optics and image processing from the Ecole Centrale Marseille, Marseille, France, in 2013.

He is currently an Associate Professor with Jiangnan University. His current research interests include image processing, computer vision, and pattern recognition.



Liang Xiao (Member, IEEE) received the B.S. degree in applied mathematics and the Ph.D. degree in computer science from Nanjing University of Science and Technology, Nanjing, China, in 1999 and 2004, respectively.

From 2009 to 2010, he was a Postdoctoral Fellow with Rensselaer Polytechnic Institute, USA. He is currently a Professor with the School of Computer Science, Nanjing University of Science and Technology. His main research interests include inverse problems in image processing, scientific computing, data mining, and pattern recognition.



Jinlong Yang received the Ph.D. degree in pattern recognition and intelligence system from Xidian University, Xi'an, China, in 2012.

He is currently an Associate Professor with Jiangnan University. His current research interests include target tracking, information fusion, and signal processing.