Deep Hyperspectral Image Fusion Network With Iterative Spatio-Spectral Regularization

Tao Huang, Weisheng Dong, *Member, IEEE*, Jinjian Wu, *Member, IEEE*, Leida Li, *Member, IEEE*, Xin Li, *Fellow, IEEE*, and Guangming Shi, *Fellow, IEEE*

***Abstract*—Physical acquisition of high-resolution hyperspectral images (HR-HSI) has remained difficult, despite its potential of resolving material-related ambiguities in vision applications. Deep hyperspectral image fusion, aiming at reconstructing an HR-HSI from a pair of low-resolution hyperspectral image (LR-HSI) and high-resolution multispectral image (HR-MSI), has become an appealing computational alternative. Existing fusion methods either rely on hand-crafted image priors or treat fusion as a nonlinear mapping problem, ignoring important physical imaging models. In this paper, we propose a novel regularization strategy to fully exploit the spatio-spectral dependency by a spatially adaptive 3D filter. Moreover, the joint exploitation of spatio-spectral regularization and physical imaging models inspires us to formulate deep hyperspectral image fusion as a differentiable optimization problem. We show how to solve this optimization problem by an end-to-end training of a model-guided unfolding network named DHIF-Net. Unlike existing works of simply concatenating spatial with spectral regularization, our approach aims at an end-to-end optimization of *iterative* spatio-spectral regularization by multistage network implementations. Our extensive experimental results on both synthetic and real datasets have shown that our DHIF-Net outperforms other competing methods in terms of both objective and subjective visual quality.**

***Index Terms*—Deep convolutional network, hyperspectral image fusion, model-guided unfolding network, spatio-spectral regularization.**

I. Introduction

WHEN compared with multi-spectral images (MSI), hyperspectral images (HSI) contain a lot more spectral bands useful to more accurately distinguish the fine characteristics of different material in the scene. Due to their rich spectral information, the class of HSI has unique advantages in many computer vision tasks - e.g., object recognition [1], detection [2], segmentation [3] and tracking [4]–[6]. However, hardware limitations often spell a curse on the trade-off between spatial and spectral resolutions. To reach high spectral resolution, practical HSI imaging systems often fail to capture scenes with high spatial resolution.

The above limitation can be overcome by computational methods known as hyperspectral image fusion [7]. Hyperspectral image fusion aims to reconstruct HR-HSI by fusing a sequence of LR-HSI with a high-resolution multispectral image (HR-MSI) [8], [9]. Given an LR-HSI $\mathbf{X}$ and an HR-MSI $\mathbf{Y}$, one can describe their relationship with the unknown HR-HSI $\mathbf{Z}$ by

$$\mathbf{Y} = \mathbf{ZR}, \mathbf{X} = \mathbf{BZ}, \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^{N\times l}$, $\mathbf{X} \in \mathbb{R}^{n\times L}$ and $\mathbf{Z} \in \mathbb{R}^{N\times L}$ ($N \gg n$ and $L \gg l$) denote the HR-MSI, the LR-HSI and the HR-HSI respectively, $N = WH$ denotes the total number of pixel of a spectral band of size $H \times W$, $\mathbf{R} \in \mathbb{R}^{L\times l}$ denotes the spectral response matrix converting a HSI to an MSI, and $\mathbf{B} \in \mathbb{R}^{n\times N}$ denotes the spatial down-sampling matrix composed of a blurring and subsampling matrices. When physical imaging models are unknown, such an inverse problem has also been named blind hyperspectral image fusion [10], [11].

Early fusion methods obtain the HR-HSI by fusing the pair of LR-HSI $\mathbf{X}$ and HR-MSI $\mathbf{Y}$ in a transform domain- e.g., principal component analysis (PCA) [12], [13] or wavelet transform [14]. The joint filtering method [15] that uses the HR-MSI as a guided image to upscale the LR-HSI has also been proposed. Non-negative matrix factorization methods have also been proposed to fuse the pair of LR-HSI and HR-MSI [16]–[18]. Other methods tackle the fusion problem by solving an optimization problem. Considering that the fusion problem is ill-posed, various regularization terms encoding the HSI prior have been proposed, including the total variation (TV) model [19] and its iterative extension [20], dictionary learning based sparsity models [21]–[24] and clustering manifold model [25]. Though promising performance has been achieved by these model-based methods, most prior models are hand-crafted, and these methods are time-consuming.

Inspired by the success of deep convolutional neural networks (DCNN) in image superresolution [26] and image denoising [27]–[29], deep learning has also been applied to the problem of hyperspectral image fusion [10], [11], [30]–[32]. In [30], a 3-D DCNN was proposed to estimate the HR-HSI

Manuscript received March 4, 2021; revised July 7, 2021 and November 7, 2021; accepted January 31, 2022. Date of publication February 22, 2022; date of current version March 4, 2022. This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0101400 and in part by the Natural Science Foundation of China under Grants 61991451, 61632019, 61621005, and 61836008. The work of Xin Li was supported in part by NSF under Grants IIS-1951504 and OAC-1940855, in part by DoJ/NIJ under Grant NIJ 2018-75-CX-0032, and in part by WV Higher Education Policy Commission Grant HEPC.dsr.18.5. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Stanley H Chan. *(Corresponding author: Weisheng Dong.)*

Tao Huang, Weisheng Dong, Jinjian Wu, Leida Li, and Guangming Shi are with the School of Artificial Intelligence, Xidian University, Xi'an, Shaanxi 710071, China (e-mail: thuang_666@stu.xidian.edu.cn; wsdong@mail.xidian.edu.cn; jinjian.wu@mail.xidian.edu.cn; ldli@xidian.edu.cn; gmshi@xidian.edu.cn).

Xin Li is with the Lane Deptartment of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506-6109 USA (e-mail: xin.li@ieee.org).

Digital Object Identifier 10.1109/TCI.2022.3152700

from a stack of the interpolated LR-HSI and the HR-MSI. The powerful learning ability of DCNN has led to state-of-the-art performance. However, by treating image fusion as a nonlinear mapping problem, existing DCNN methods have ignored the important information contained in the observation models. To address this issue, physical imaging models have been taken into account in recent works [10], [11], [31] leading to further improvement.

In this paper, we propose an optimization-inspired HSI fusion network named DHIF-Net. To fully exploit the spatio-spectral dependency of HSIs, we propose a novel spatio-spectral regularization, where we enforce each pixel can be well predicted by using a spatially varying 3D filter to filter its spatial-spectral neighbors. A deep neural network is proposed to estimate the spatially varying 3D filters from the pair of LR-HSI and HR-MSI. With the guidance of HR-MSI, we can accurately estimate the intermediate 3D filters for each pixel in the HR-HSI. Based on the proposed spatio-spectral regularization, the HSI fusion objective function is differentiable and thus can be solved by end-to-end training. Unlike previous work simply concatenating spatial with spectral regularization [32], our approach aims at an end-to-end optimization of *iterative* spatio-spectral regularization by multistage network implementations. Extensive experimental results on both synthetic and real-world datasets have shown that the proposed DHIF-Net advances the state-of-the-art in deep hyperspectral image fusion.

II. Related Works

In this section, we briefly review the conventional model-based fusion methods and the recently proposed deep learning-based fusion methods.

A. Model-Based Methods

Traditional methods fuse an HR-MSI and an LR-HSI in transform domains [12]–[14], where certain components of the LR-HSI are replaced or fused with those of HR-MSI. Although these methods can increase the spatial resolution, they often introduce spectral distortions. In recent years, reconstruction-based methods have been proposed, which formulate the fusion problem as an optimization problem. As the reconstruction from the pair of LR-HSI and HR-MSI is an ill-posed problem, various regularization terms have been proposed [19], [21]–[23]. In [19] total variation (TV) regularization has been adopted to suppress the artifacts of the recovered HSIs. Dictionary learning-based sparse models, which represent each spectral pixel as a linear combination of a few reflectance spectral bases, have been proposed as prior models to regularize the fusion process [21]–[23]. Structured sparsity model considering the spatial correlations between the pixels has also been proposed to further improve the fusion performance [24]. The clustering manifold model that represents each spectral pixel as a combination of its neighboring pixels has also been exploited as a prior for HSI [25]. By exploiting the prior knowledge of HSI, the iterative model-based methods have achieved very promising experimental results. The major drawbacks of these methods are that the prior is hand-crafted and the parameters of these models cannot be jointly optimized.

B. Deep Learning-Based Methods

Recently, deep learning-based hyperspectral image fusion methods have attracted increasing attentions [10], [11], [30]–[32]. Deep pansharpening methods that take pairs of HR-MSIs and LR-HSIs as input and produce the corresponding HR-HSIs with specifically designed DCNNs. These methods consider the fusion as a highly nonlinear mapping problem and thereby ignore the observation models of HSI. To overcome this drawback, in [31] and [11] a DCNN denoiser was proposed as a regularization for estimating an intermediate result of the unknown HR-HSI, which was further used to solve for the final HR-HSI with a low-rank constraint. More recently, a deep neural network [10] was built based on the back-projection framework, where a DCNN denoiser was used to suppress the projection error to refine the estimate of HR-HSI. Although these methods have shown better performance, the domain knowledge of HSI has not been fully exploited. In particular, the rich dynamics of spatio-spectral dependency calls for a more principled solution to the design of spatio-spectral regularization strategy, which sets up the stage for this work.

III. DHIF-Net for Hyperspectral Image Fusion

As the recovery of HR-HSI $\mathbf{Z} \in \mathbb{R}^{N\times L}$ from the pair of HR-MSI $\mathbf{Y} \in \mathbb{R}^{N\times l}$ and LR-HSI $\mathbf{X} \in \mathbb{R}^{n\times L}$ is an ill-posed problem, the estimation of HR-HSI $\mathbf{Z}$ is often obtained by solving the following regularized Least-Squares problem

$$\mathbf{Z} = \underset{\mathbf{Z}}{\operatorname{argmin}} ||\mathbf{Y} - \mathbf{ZR}||_F^2 + ||\mathbf{X} - \mathbf{BZ}||_F^2 + \lambda J(\mathbf{Z}), \quad (2)$$

where $J(\mathbf{Z})$ is the regularization term encoding the prior knowledge of $\mathbf{Z}$. To improve the estimation performance, various regularization techniques including TV-based [19] and sparsity-based [21]–[23] have been proposed. However, most existing regularization techniques are designed for photographic images, which cannot fully exploit the spatio-spectral dependency among 3D observation data. In this paper, we propose a novel 3D spatio-spectral regularization for better characterizing the class of HSIs. Inspired by the autoregressive model for HSI [23], we propose to predict each pixel of $\mathbf{Z}$ as a weighted average of its surrounding pixels - i.e.,

$$z_{i,s} = \sum_a \sum_b \boldsymbol{k}_{i,s}(a,b)\mathbf{Z}_{i,s}(a,b) + e_{i,s}, \quad (3)$$

where $\boldsymbol{k}_{i,s} \in \mathbb{R}^{m^2\times m}$ is the 3D filter of size $m \times m \times m$ for pixel $z_{i,s}$, $\mathbf{Z}_{i,s} \in \mathbb{R}^{m^2\times m}$ denotes the local spatio-spectral neighborhood of the pixel $z_{i,s}$, a and b denote the spatial and spectral indexes of the filters $\boldsymbol{k}_{i,s}$, respectively, and $e_{i,s}$ denotes the prediction error of pixel $z_{i,s}$. Such 3D filter-based spatio-spectral regularization is more powerful than the existing spatial or spectral regularization for HSI prior modeling.

Different from previous work [23], we propose to estimate $\boldsymbol{k}_{i,s}$ by a DCNN from the guidance of HR-MSI $\mathbf{Y}$ and LR-HSI $\mathbf{X}$ (more details will be described in the next section). With

Algorithm 1: Proposed Iterative HSI-Fusion Algorithm.

- **Input**: The HR-MSI $\mathbf{Y}$ and the LR-HSI $\mathbf{X}$
- **Initialization**:
  (1) Set parameters λ, δ, and $t = 0$;
  (2) Initialize $\mathbf{Z}^{(0)}$ with bicubic interpolation.
- **While** does not converge **do**
  (1) Estimate the filtering matrix $\mathbf{K}^{(t)}$;
  (2) Compute the filtered result $\mathbf{U}^{(t)}$ using $\mathbf{K}^{(t)}$ and $\mathbf{Z}^{(t)}$;
  (3) Update $\mathbf{Z}^{(t)}$ via (6);
  (4) $t = t + 1$.

End While

- **Output**: The HR-HSI $\mathbf{Z}$

the proposed spatio-spectral regularization, the global objective function can be expressed as

$$\mathbf{Z} = \underset{\mathbf{Z}}{\operatorname{argmin}} ||\mathbf{Y} - \mathbf{Z}\mathbf{R}||_F^2 + ||\mathbf{X} - \mathbf{B}\mathbf{Z}||_F^2 + \lambda||\boldsymbol{z} - \mathbf{K}\boldsymbol{z}||_2^2, \tag{4}$$

where we enforce that the estimated $\mathbf{Z}$ should be close to its filtered version, which significantly reduces the solution space. Note that $\mathbf{K} \in \mathbb{R}^{NL \times NL}$ denotes the filtering matrix constructed from the 3D filters $\boldsymbol{k}_{i,s}$, and $\boldsymbol{z} = \text{vec}(\mathbf{Z}) \in \mathbb{R}^{NL \times 1}$, where $\text{vec}(\cdot)$ denotes the vectorization operator.

The above objective function can be solved by standard gradient descent algorithms. However, most gradient descent algorithms involve the matrix multiplication with the transposed $\mathbf{K}^\top$, which is difficult to explicitly construct and compute. To overcome this difficulty, we propose to iteratively solve the following surrogate objective function

$$\mathbf{Z}^{(t+1)} = \underset{\mathbf{Z}}{\operatorname{argmin}} ||\mathbf{Y} - \mathbf{Z}\mathbf{R}||_F^2 + ||\mathbf{X} - \mathbf{B}\mathbf{Z}||_F^2 + \lambda||\boldsymbol{z} - \mathbf{K}^{(t)}\boldsymbol{z}^{(t)}||_2^2, \tag{5}$$

where the current estimate $\mathbf{Z}^{(t)}$ is used to compute the spatio-spectral regularization term, and the filtering matrix $\mathbf{K}^{(t)}$ is estimated from the guided HR-MSI $\mathbf{Y}$ and the current estimate $\mathbf{Z}^{(t)}$. The above objective function is quadratic and can be solved in closed form. But still it requires the computation of the inverse of a large matrix, which can be slow. Instead, we propose to solve it via a single-step gradient descent by

$$\mathbf{Z}^{(t+1)} = \mathbf{Z}^{(t)} - 2\delta\{(\mathbf{Z}^{(t)}\mathbf{R} - \mathbf{Y})\mathbf{R}^\top + \mathbf{B}^\top(\mathbf{B}\mathbf{Z}^{(t)} - \mathbf{X}) + \lambda(\mathbf{Z}^{(t)} - \mathbf{U}^{(t)})\}, \tag{6}$$

where $\mathbf{U}^{(t)} = \text{reshape}(\mathbf{K}^{(t)}\boldsymbol{z}^{(t)}) \in \mathbb{R}^{N \times L}$, where $\text{reshape}(\cdot)$ denotes an operator representing the filtered version of $\mathbf{Z}$ with the estimated 3D filters, and δ is the step size. The proposed iterative HSI-Fusion algorithm with the newly constructed spatio-spectral regularization is summarized in the following **Algorithm 1**.

Regarding the estimation of 3D filters, one may use the guided filtering [33], [34] or nonlocal means filtering methods [35], [36] to estimate the spatially variant filters. A common weakness of these conventional methods is that they cannot optimize the filters jointly with other parameters in an end-to-end manner. Inspired by [37] and [38] which proposed a U-net to predict kernels in an end-to-end manner, we also used U-net as the main structure to estimate the filters from the guided HR-MSI $\mathbf{Y}$ and the current estimate $\mathbf{Z}^{(t)}$. Further, inspired by recent work [29], we propose to unfold Algorithm 1 into a DCNN-based implementation. Such model-guided network design enjoys not only transparency (interpreted as a numerical solution to the well-defined optimization problem) but also flexibility (all hyperparameters including spatio-spectral regularization related and physical-imaging model related can be tuned by end-to-end optimization).

IV. DHIF-Net: Deep Network Implementation

A. The Overall Network Architecture

The overall network architecture of the proposed DHIF-Net is shown in Fig. 1(a), which contains T stages corresponding to totally T iterations of **Algorithm 1**. As shown in Fig. 1(a), the current estimate $\mathbf{Z}^{(t)}$ and the inputs are fed into the reconstruction module to minimize the reconstruction error as

$$\mathbf{Z}^{(t+1/2)} = \mathbf{Z}^{(t)} - 2\delta\{(\mathbf{Z}^{(t)}\mathbf{R} - \mathbf{Y})\mathbf{R}^\top + \mathbf{B}^\top(\mathbf{B}\mathbf{Z}^{(t)} - \mathbf{X})\}. \tag{7}$$

The output of the reconstruction module $\mathbf{Z}^{(t+1/2)}$ is then added to the output of the spatio-spectral regularization module as

$$\mathbf{Z}^{(t+1)} = \mathbf{Z}^{(t+1/2)} - 2\delta\lambda(\mathbf{Z}^{(t)} - \mathbf{U}^{(t)}). \tag{8}$$

We note that similar unfolding strategies originated from the strategy of iterative regularization in model-based image restoration [20], which has been proposed for image super-resolution [39], [40] and denoising [29], [41] in the literature. However, such an iterative spatio-spectral regularization method has not been fully considered for HSI-related restoration problems yet. Different from previous methods (mostly in the spatial domain), the newly constructed spatio-spectral regularization calls for special attention in the spatial spectral domain. Unlike existing works of simply concatenating spatial with spectral regularization (e.g., [32]), our approach is capable of optimizing *iterative* spatio-spectral regularization by multistage network implementations in an end-to-end manner.

B. The Spatio-Spectral Regularization Module

As shown in Fig. 1(b), the current estimate $\mathbf{Z}^{(t)}$ and the guided HR-MSI $\mathbf{Y}$ are concatenated as the input to a U-net for extracting salient features from the inputs. The extracted features are then fed into a filter generator to generate spatially adaptive and content-aware 3D filters (a computationally efficient separable implementation of 3D filters will be elaborated next). These spatially adaptive 3D filters have advantages in capturing the rich dynamics of spatio-spectral dependency and will be updated with the update of $\mathbf{Z}$ at each stage. However, generating these spatially adaptive 3D filters requires a large amount of GPU memory and computational resources, which is not affordable, especially for a multistage implementation. To reduce memory consumption and improve computational efficiency simultaneously, we propose to factorize each 3D filter into three separable

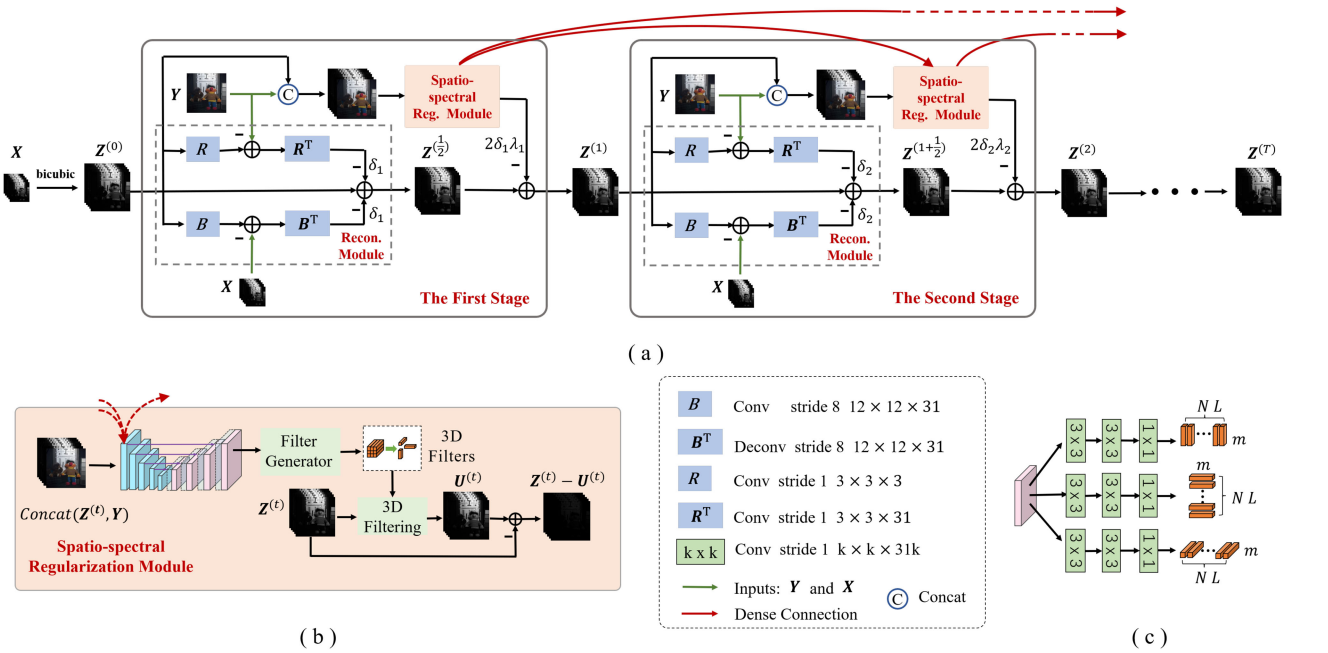


Fig. 1. Architecture of the proposed network for hyperspectral image fusion. The architecture of (a) The overall network; (b) The spatio-spectral regularization module; (c) The 3D filter generator.

1D filters- as

$$\mathcal{K}_{i,s} = \boldsymbol{r}_{i,s} \otimes \boldsymbol{c}_{i,s} \otimes \boldsymbol{s}_{i,s}, \tag{9}$$

where $\mathcal{K}_{i,s} \in \mathbb{R}^{m\times m\times m}$ denotes the 3D filter coefficient array, $\boldsymbol{r}_{i,s} \in \mathbb{R}^m$, $\boldsymbol{c}_{i,s} \in \mathbb{R}^m$ and $\boldsymbol{s}_{i,s} \in \mathbb{R}^m$ denote the learned corresponding row, column and spectral dimension 1D filters, and $\otimes$ denotes the tensor product. The convolution of the separable 3D filters with local neighboring pixels can be performed by first convoluting the local HSI cube $\mathbf{Z}_{i,s}$ with $\boldsymbol{r}_{i,s}$ along the row dimension, and then along the column dimension with $\boldsymbol{c}_{i,s}$, followed by convoluting with $\boldsymbol{s}_{i,s}$ along the spectral dimension. As shown in Fig. 1(c), we have developed three branches for learning three separable 1D filters, respectively.

For the kernels of size $m \times m \times m$, the computation complexity of (3) is $O(m^3)$, but the computation complexities of factorized 3D kernels reduce to $O(3 \cdot m)$. Moreover, while the size of the 3D filters in the original manner is $O(N \cdot L \cdot m^3)$, the total size of the three filters in the factorization manner reduce to $O(3 \cdot N \cdot L \cdot m)$. For the CAVE dataset, the proposed factorization strategy can save nearly 16 times memory theoretically, where $W = H = 512$, $B = 31$, and $m = 7$. Note that as shown in Fig. 1(b), we update the spatially variant filters for each stage based on the improved estimate $\mathbf{Z}^{(t)}$ of the HR-HSI, which further improves the HSI recovery performance. The U-net contains five encoding blocks and four decoding blocks. Each block has two convolutional layers with 3×3 kernels and two ReLU layers. Four average pooling layers followed by the first four encoding blocks reduce the feature resolution with a stride of two, and four bilinear interpolation operations with a scaling factor of two are employed to increase the input feature resolution of the four decoding blocks. A total of nine blocks (i.e., five encoding blocks and four decoding blocks) have 64, 128, 256, 512, 512, 512, 256, 128, 128 output channels, respectively. Additionally, we have introduced dense connections to bridge the first feature maps of U-net from the previous stages to the subsequent stages. These interstage dense connections can effectively alleviate the notorious vanishing gradient problem by facilitating the information flow across the stages. The benefit of introducing dense connections has been experimentally justified (please refer to see Section V-B).

C. Reconstruction Module

As shown in the dotted block of Fig. 1(a), the reconstruction module exactly executes (7). Note that the pair of dual operations $(\mathbf{R}, \mathbf{R}^T)$ and $(\mathbf{B}, \mathbf{B}^T)$ in the upper and lower channels correspond to the second and third terms on the right side of (7) respectively. The spatial degradation operator $\mathbf{B}$ representing spatial blur and down-sampling hybrid operators can be performed by a convolutional layer with the stride s. Inversely, the transposed version $\mathbf{B}^\top$ can be performed by a deconvolutional layer with the same stride s. The spectral degradation operator $\mathbf{R}$ represents the spectral response function of the multispectral imaging sensor and converts the HR-HSI into the HR-MSI. Thus, $\mathbf{R}$ and its transposed version $\mathbf{R}^\top$ can be easily modeled by a convolutional layer, respectively. Similar to [10] all these operators implemented on convolutional layers are followed by a ReLU nonlinear function.

D. Network Training

Through end-to-end training, the network parameters $\boldsymbol{\Theta}$ and the parameters δ and λ can be learned jointly. The learnable network parameters $\boldsymbol{\Theta}$ include the parameters of U-Net, reconstruction module and filter generator. δ and λ in different stages are not shared and empirically initialized as 0.1 and 1. Except δ

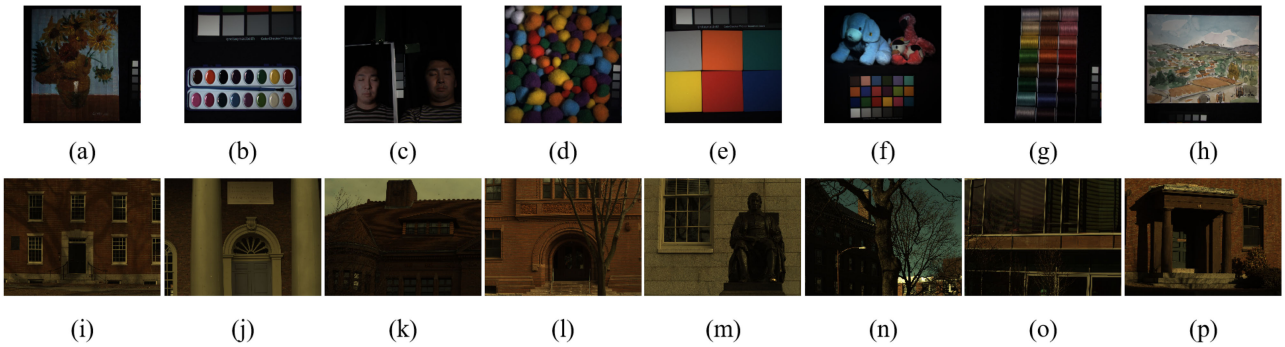


Fig. 2. HR RGB images from the CAVE (the first row) [42] and Harvard (the second row) [43] datasets. (a) *oil_painting*; (b) *paints*; (c) *photo_and_face*; (d) *pompoms*; (e) *sponges*; (f) *stuffed_toys*; (g) *thread_spools*; (h) *watercolors*; (i) *imge3*; (j) *imge4*; (k) *imge5*; (l) *imge6*; (m) *imge7*; (n) *imgf2*; (o) *imgf3*; (p) *imgf7*.

and λ, all stages share the same network parameters. The l_1 loss function is adopted to train the proposed network, and the loss function can be written as

$$(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\delta}}) = \underset{\boldsymbol{\Theta}, \boldsymbol{\lambda}, \boldsymbol{\delta}}{\operatorname{argmin}} \frac{1}{D} \sum_{d=1}^{D} ||\mathcal{F}(\mathbf{Y}_d, \mathbf{X}_d; \boldsymbol{\Theta}, \boldsymbol{\lambda}, \boldsymbol{\delta}) - \mathbf{Z}_d||_1, \tag{10}$$

where D denotes the total number of the training samples, $\mathbf{Z}_d$, $\mathbf{Y}_d$ and $\mathbf{X}_d$ represent the d^{th} pair of the label HR-HSI, HR-MSI and LR-HSI, respectively, $\mathcal{F}(\cdot)$ denotes the output of the proposed network given inputs and the parameters. We used the ADAM optimizer [44] to train the proposed network by setting $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. The learning rate is set to 10^{-4}. The Xavier initializers [45] were used to initialize the parameters of the convolutional layers. We have implemented the proposed method using PyTorch and train network using a single Nvidia Titan XP GPU.

V. Experimental Results

A. Experimental Setup

To verify the performance of the proposed HSI fusion method, extensive experiments on two public HSI datasets (i.e., the Harvard dataset [43] and the CAVE dataset [42]) have been conducted. The Harvard dataset consists of 50 indoor and outdoor HSIs under daylight, and each HSI has 31 spectral bands of size 1040×1392. The CAVE dataset consists of 32 HSIs of size $512 \times 512 \times 31$. Similar to [10], the first 30 and 20 HSIs are used for training, and the remaining 20 and 12 HSIs for testing on the Harvard dataset and the CAVE dataset, respectively. Some simulated RGB images used for testing on the Harvard and CAVE datasets are shown in Fig. 2. We have used the same degradation operator to simulate the HR-MSIs and the LR-HSIs for both datasets. The original HSIs are used as the ground truth for objective performance evaluation. To simulate the LR-HSIs, we first apply an $s \times s$ ($s = 8, 16, 32$) Gaussian filter with a standard deviation of two to blur the HR-HSIs, and then downsample the blurred HSIs in both the horizontal and vertical directions with a scaling factor of eight. Similar to [24] and [10], the HR RGB images are generated by applying the spectral response $\mathbf{R}$ of a Nikon D700 camera[1]. We randomly extract $96 \times 96 \times 31$ overlapped patches from the HR-HSIs and then use these patches to simulate corresponding LR-HSIs of size $\frac{96}{s} \times \frac{96}{s} \times 31$ and HR-MSIs of size $96 \times 96 \times 3$ for training. Random flipping and rotation are used for image argumentation.

TABLE I
Ablation Study on the Effects of the Number of Stages for the CAVE Dataset and Scaling Factor 8

Dataset	CAVE					
Model	DHIF-Net					
T	1	2	3	4	5	6
PSNR	49.27	49.59	49.74	49.87	49.79	49.82
SAM	3.12	3.08	3.07	3.02	2.98	3.00
ERGAS	0.53	0.52	0.51	0.51	0.51	0.51
SSIM	0.995	0.995	0.995	0.995	0.995	0.995
Model	DHIF-Net-Simpler					
T	1	3	5	7	9	11
PSNR	48.21	48.73	48.93	49.41	49.34	48.78
SAM	3.19	3.21	3.10	3.02	3.07	3.11
ERGAS	0.60	0.58	0.55	0.52	0.53	0.55
SSIM	0.994	0.994	0.995	0.995	0.995	0.995

Four quality metrics are employed to evaluate the performance of the HSI SR method, including the peak-signal-to-noise (PSNR), spectral angle mapper (SAM) [50], relative dimensionless global error in synthesis (ERGAS) [51], and the structural similarity index (SSIM) [52].

B. Ablation Studies

To verify the impact of the number of stages T, the dense connections between the U-nets in different stages, and the filter size m on the fusion performance, we have conducted several ablation studies. Table I shows the results of the proposed method with different number of stages. From Table I one can see that the performance improves with the increase of T. However, the reconstructed results become worse when $T > 4$. We conjecture that deeper networks are generally more difficult to train because of the notorious gradient vanishing problem, resulting in degraded results. To achieve a good trade-off between fusion performance and computational complexity, we have set $T = 4$

[1] Available at: https://www.maxmax.com/spectral} response.htm

TABLE II
ABLATION STUDY ON THE EFFECTS OF DENSE CONNECTIONS FOR THE CAVE DATASET AND SCALING FACTOR 8

Dataset	Dense Connection	PSNR	SAM	ERGAS	SSIM
CAVE	✗	49.37	3.06	0.52	0.995
	✓	49.87	3.02	0.51	0.995

TABLE III
ABLATION STUDY ON THE EFFECTS OF THE FILTER SIZES FOR THE CAVE DATASET AND SCALING FACTOR 8

m	3	5	7	9	11
PSNR	49.38	49.54	49.79	49.80	49.87
SAM	3.11	3.04	3.01	3.01	3.02
ERGAS	0.52	0.51	0.51	0.50	0.51
SSIM	0.995	0.995	0.995	0.995	0.995

TABLE IV
COMPARISON RESULTS OF THE PROPOSED SEPARABLE FILTERS AND THE TRADITIONAL 3D FILTERS

Dataset	CAVE		
Filters	$3 \times 3 \times 3$ (traditional)	3 + 3 + 3 (separable)	9 + 9 + 9 (separable)
PSNR	49.44	49.38	49.80
SAM	3.02	3.11	3.01
ERGAS	0.53	0.52	0.50
SSIM	0.995	0.995	0.995

in our implementation. To verify our conjecture, we have conducted more experiments by using a simpler network to replace the U-Net at each stage. The simpler network is a lightweight U-Net that contains four encoding blocks and three decoding blocks. The feature map of each block has only 64 output channels. This variant is denoted as DHIF-Net-Simpler. From Table I, the performance of DHIF-Net-Simpler also improves with the increase of T when $T \leq 7$. Due to the simpler network, DHIF-Net-Simpler uses more stages than DHIF-Net to obtain the best results. Similarly, the reconstructed results of DHIF-Net-Simpler become worse when T is greater than the best T, i.e., $T > 7$. When $T = 11$, the network is very deep, resulting in degraded results. The experiments of DHIF-Net-Simpler can verify our conjecture further. The comparison results of the proposed methods with and without dense connections are shown in Table II. From Table II, we can observe that dense connections improve the HSI fusion results by about 0.5 dB. The results of the proposed methods with different filter sizes are shown in Table III. One can see that 3D filters of larger size indeed improve the overall HSI fusion performance. The performance improvement saturates at $m = 7$. Therefore, we set $m = 7$ in our implementation.

We have conducted the comparison between the proposed separable filters and the traditional filters. The comparison results are shown in Table IV. We only compared the 3D filters of size 3. Because when $m > 3$, the traditional filters require a large amount of GPU memory and computational resources. It's beyond that we can afford. Compared with the traditional filters, the proposed separable filters have lost 0.06 dB in PSNR. With the same computation complexity, the proposed separable filter of size 9 outperforms the traditional filters of size 3 by up to 0.36 dB. It can be seen that a larger receptive field can improve the fusion results. We want to note that the main purpose of factorizing each 3D filter into three separable 1D filters is to reduce memory consumption and improve computational efficiency simultaneously.

To further explore where the improvements arise from, we have conducted more experiments to discuss the technical contributions of the proposed methods. First, we have realized **Algorithm 1** by vanilla convex optimization, i.e., the conventional guided filtering method, where we used the entire HR-MSI as a guide to estimate the filters. After obtaining the filtering matrix $\mathbf{K}^{(t)}$, we utilized filtering operation to compute the filtered result $\mathbf{U}^{(t)}$ and updated $\mathbf{Z}^{(t)}$ via (6). The hyperparameters δ and λ were set to 0.9 and 0.001, respectively. In our implementation, **Algorithm 1** typically converges after 25 iterations. In addition, for comparing with static 3D filters, a U-net is developed to extract salient features and a 3D convolutional layer followed by the U-net estimates $\mathbf{U}$ directly. The developed network has similar parameters to the proposed method. All parameters are trained in an end-to-end manner. After training, the 3D filters of the 3D convolutional layer are static and fixed for testing. Table V shows the comparison results of vanilla convex optimization, optimization with static 3D filter, and optimization with dynamic 3D filter. From Table V, we can see that the two deep learning implementations dramatically outperform vanilla convex optimization (i.e., the conventional guided filtering method). Implemented with the same number of stages, optimization with dynamic 3D filters outperforms optimization with static 3D filters. Finally, unrolling can improve the fusion results. In summary, the proposed DHIF-Net combines the powerful learning capability of DCNN with the rich dynamics of spatio-spectral dependency and multistage network implementations, thus leading to substantial performance improvement.

In addition, we have also studied the single spectral super-resolution problem. When inputting only LR-HSI and using one stage, the single spectral super-resolution result is very bad (PSNR = 32.39 dB, SAM = 5.58, ERGAS = 3.34, SSIM = 0.897) compared with the fusion result (PSNR = 49.27 dB, SAM = 3.12, ERGAS = 0.53, SSIM = 0.995). The reason is that the LR-HSI has less high-frequency spatial details, resulting in estimating more inaccurate spatially varying 3D filters. However, in the fusion problem, the estimated spatially varying 3D filters are more accurate due to the guidance of the HR-HSI.

C. Comparison With State-of-the-Art Fusion Methods

We have compared the proposed HSI fusion method with several state-of-the-art methods, including four model-based HSI fusion methods (i.e., the CSU method [46], the HySure method [47], the NSSR method [24] and the CSTF method [48]) and three recently proposed deep learning based methods (i.e., the DHSIS method [49], the MHF-net method [31] and the DBIN method [10]). The source codes of other methods used

TABLE V
AVERAGE PSNR, SAM, ERGAS, AND SSIM RESULTS ON THE CAVE DATASET AND THE HARVARD DATASET FOR DISCUSSING THE TECHNICAL CONTRIBUTIONS

Dataset	CAVE							
Method	convex optimization	static 3D filters				dynamic 3D filters			
iteration / stage	25	1	2	3	4	1	2	3	4
PSNR	43.81	48.78	49.14	49.24	49.37	49.27	49.59	49.74	49.87
SAM	4.33	3.24	3.22	3.16	3.16	3.12	3.08	3.07	3.02
ERGAS	0.87	0.55	0.53	0.53	0.53	0.53	0.52	0.51	0.51
SSIM	0.989	0.994	0.994	0.995	0.995	0.995	0.995	0.995	0.995

TABLE VI
AVERAGE PSNR, SAM, ERGAS, AND SSIM RESULTS OF THE TEST METHODS ON THE CAVE DATASET AND THE HARVARD DATASET FOR GAUSSIAN BLUR KERNEL AND SCALING FACTORS 8/16/32

Factor	8							
Dataset	CAVE				Harvard			
Method	PSNR	SAM	ERGAS	SSIM	PSNR	SAM	ERGAS	SSIM
CSU [46]	41.23	6.58	1.15	0.982	45.41	3.88	1.39	0.984
HySure [47]	37.04	11.19	1.85	0.960	42.02	4.67	1.79	0.977
CSTF [48]	42.34	6.48	0.98	0.975	42.24	5.16	1.62	0.961
NSSR [24]	44.07	4.40	0.83	0.987	46.08	3.68	1.28	0.984
DHSIS [49]	46.48	3.89	0.66	0.992	46.53	3.57	1.27	0.985
MHF-net [31]	46.46	4.37	0.67	0.992	46.89	3.61	1.27	0.985
DBIN [10]	48.73	3.11	0.55	0.994	47.39	3.47	1.19	0.985
DHIF-Net (ours)	**49.79**	**3.01**	**0.51**	**0.995**	**47.55**	**3.40**	**1.14**	**0.986**
Factor	16							
Dataset	CAVE				Harvard			
Method	PSNR	SAM	ERGAS	SSIM	PSNR	SAM	ERGAS	SSIM
CSU [46]	38.73	8.58	0.76	0.970	44.18	4.23	0.77	0.982
HySure [47]	32.17	17.55	1.59	0.908	37.92	5.75	1.28	0.959
CSTF [48]	40.59	7.57	0.59	0.971	42.94	5.41	0.83	0.961
NSSR [24]	39.62	6.46	0.68	0.979	44.44	3.98	0.79	0.982
DHSIS [49]	39.71	6.00	0.74	0.976	45.63	3.88	0.70	0.984
MHF-net [31]	44.56	4.76	0.40	0.990	46.24	3.73	0.64	0.984
DBIN [10]	44.15	4.00	0.43	0.992	46.38	3.61	0.63	**0.985**
DHIF-Net (ours)	**46.46**	**3.66**	**0.34**	**0.993**	**46.61**	**3.60**	**0.60**	**0.985**
Factor	32							
Dataset	CAVE				Harvard			
Method	PSNR	SAM	ERGAS	SSIM	PSNR	SAM	ERGAS	SSIM
CSU [46]	36.52	9.61	0.48	0.959	42.27	4.74	0.43	0.978
HySure [47]	26.62	23.36	1.60	0.822	35.10	8.04	0.82	0.931
CSTF [48]	38.99	9.08	0.34	0.968	41.72	5.39	0.43	0.962
NSSR [24]	36.78	10.46	0.46	0.974	43.08	4.39	0.47	0.982
DHSIS [49]	38.36	8.45	0.45	0.963	44.44	4.29	0.38	0.983
MHF-net [31]	42.17	6.03	0.26	0.986	45.26	4.00	0.34	0.984
DBIN [10]	40.72	**4.86**	0.33	**0.989**	44.91	3.94	0.34	0.984
DHIF-Net (ours)	**42.83**	4.89	**0.24**	**0.989**	**45.55**	**3.80**	**0.31**	**0.985**

in this comparison study are released by the authors. For the reason of fairness, we have retrained other deep learning-based methods on the same training dataset. The average results of the test methods on the CAVE dataset and the Harvard dataset are shown in Table VI. From Table VI, we can see that the deep learning-based methods dramatically outperform conventional model-based methods. The proposed DHIF-Net outperforms other deep learning-based methods. The PSNRs gain over the second-best method in this comparison group can be up to 1.06 dB, 1.90 dB, and 0.66 dB on the CAVE dataset for scaling factors 8, 16, and 32, respectively. As the Harvard dataset is less challenging, the improvement over other methods on this dataset is smaller. Figs. 3, 4, 5, and 6 show the visual quality comparison results and the reconstructed spectra of the selected patches of

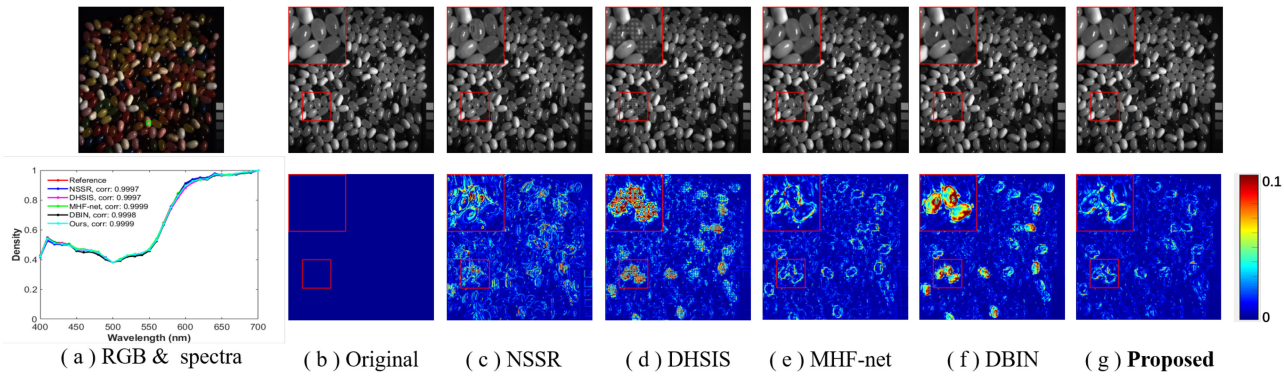


Fig. 3. Reconstructed images of $jelly\_beans$ in CAVE dataset at 700 nm with Gaussian blur kernel and scaling factor s = 8. The first row shows the reconstructed images, and the second row shows the error images of the competing methods. (a) The RGB image and the reconstructed spectra of the selected patch (indicated by a green box); (b) The ground truth images; (c) The NSSR method [24] (PSNR = 39.49 dB, SAM = 4.00, ERGAS = 1.10, SSIM = 0.984); (d) The DHSIS method [49] (PSNR = 41.98 dB, SAM = 3.88, ERGAS = 0.79, SSIM = 0.987); (e) The MHF-net method [31] (PSNR = 44.04 dB, SAM = 3.58, ERGAS = 0.64, SSIM = 0.992); (f) The DBIN method [10] (PSNR = 43.76 dB, SAM = 3.00, ERGAS = 0.60, SSIM = 0.993); (g) The proposed **DHIF-Net** method (PSNR = **46.09** dB, SAM = **2.66**, ERGAS = **0.52**, SSIM = **0.995**).

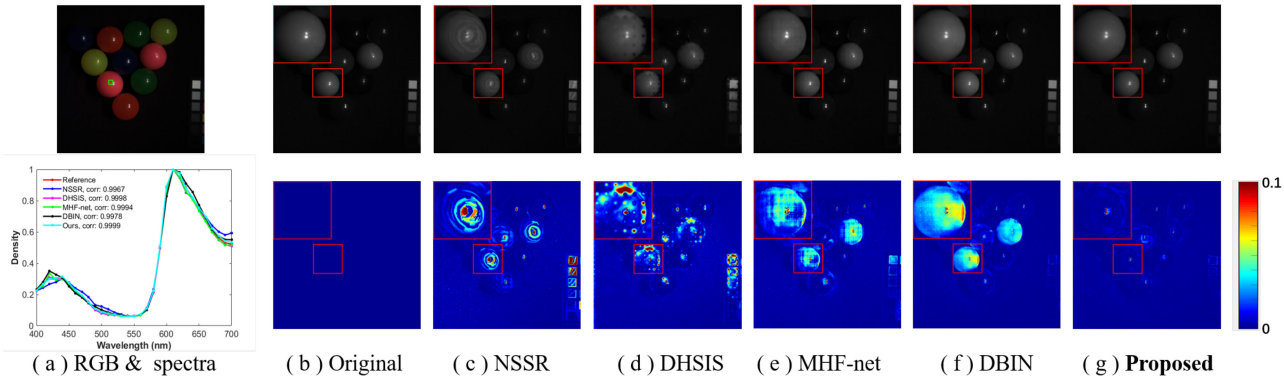


Fig. 4. Reconstructed images of $superballs$ in CAVE dataset at 420 nm with Gaussian blur kernel and scaling factor s = 16. The first row shows the reconstructed images, and the second row shows the error images of the competing methods. (a) The RGB image and the reconstructed spectra of the selected patch (indicated by a green box); (b) The ground truth images; (c) The NSSR method [24] (PSNR = 40.58 dB, SAM = 7.58, ERGAS = 1.16, SSIM = 0.981); (d) The DHSIS method [49] (PSNR = 41.49 dB, SAM = 6.98, ERGAS = 1.04, SSIM = 0.981); (e) The MHF-net method [31] (PSNR = 46.52 dB, SAM = 5.87, ERGAS = 0.64, SSIM = 0.990); (f) The DBIN method [10] (PSNR = 44.46 dB, SAM = 4.57, ERGAS = 0.75, SSIM = 0.994); (g) The proposed **DHIF-Net** method (PSNR = **49.65** dB, SAM = **4.17**, ERGAS = **0.45**, SSIM = **0.995**).

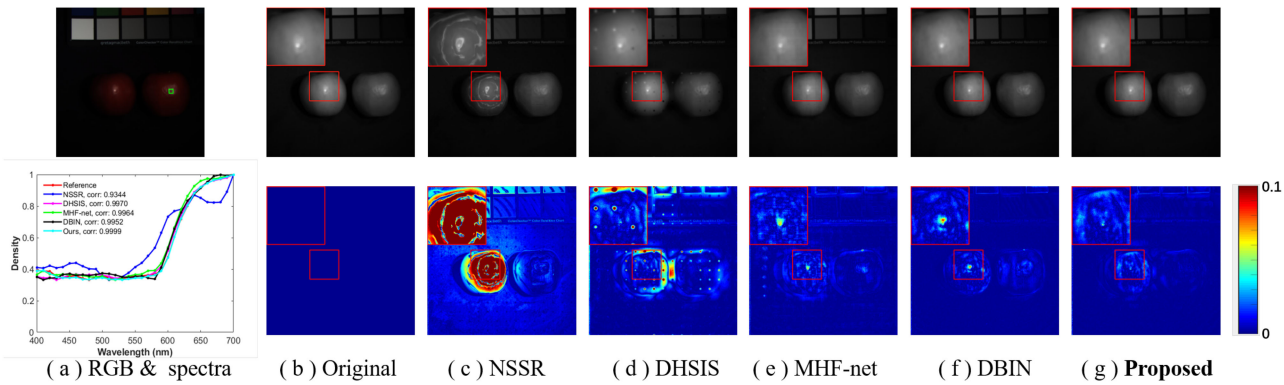


Fig. 5. Reconstructed images of $real\_and\_fake\_apples$ in CAVE dataset at 700 nm with Gaussian blur kernel and scaling factor s = 32. The first row shows the reconstructed images, and the second row shows the error images of the competing methods. (a) The RGB image and the reconstructed spectra of the selected patch (indicated by a green box); (b) The ground truth images; (c) The NSSR method [24] (PSNR = 38.93 dB, SAM = 12.53, ERGAS = 0.70, SSIM = 0.973); (d) The DHSIS method [49] (PSNR = 43.45 dB, SAM = 7.69, ERGAS = 0.58, SSIM = 0.977); (e) The MHF-net method [31] (PSNR = 50.56 dB, SAM = 6.10, ERGAS = 0.23, SSIM = 0.991); (f) The DBIN method [10] (PSNR = 53.33 dB, SAM = **4.06**, ERGAS = **0.17**, SSIM = **0.997**); (g) The proposed **DHIF-Net** method (PSNR = **53.44** dB, SAM = 4.25, ERGAS = **0.17**, SSIM = 0.996).

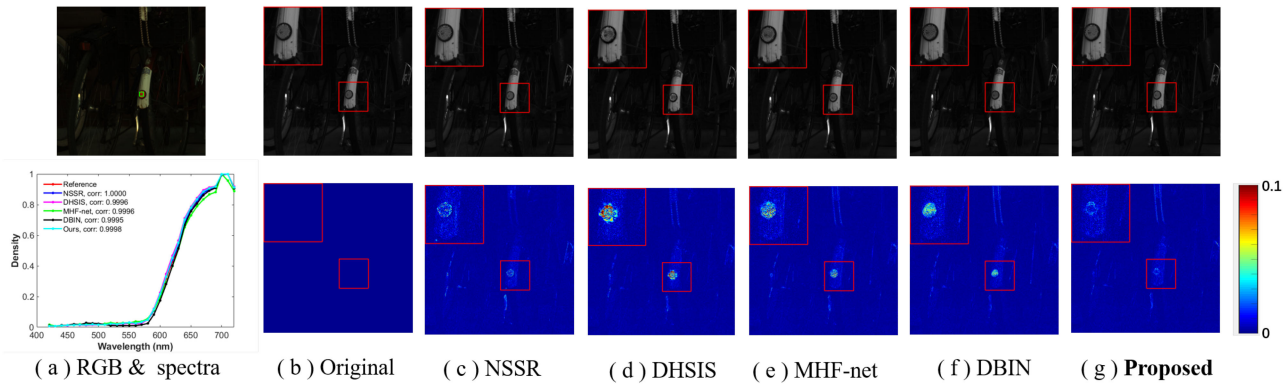


Fig. 6. Reconstructed images of *imgf5* in HARVARD dataset at 610 nm with Gaussian blur kernel and scaling factor s = 8. The first row shows the reconstructed images, and the second row shows the error images of the competing methods. (a) The RGB image and the reconstructed spectra of the selected patch (indicated by a green box); (b) The ground truth images; (c) The NSSR method [24] (PSNR = 50.68 dB, SAM = 3.58, ERGAS = 1.27, SSIM = 0.993); (d) The DHSIS method [49] (PSNR = 50.95 dB, SAM = 3.50, ERGAS = 1.19, SSIM = 0.993); (e) The MHF-net method [31] (PSNR = 51.89 dB, SAM = 3.52, ERGAS = 1.16, SSIM = 0.993); (f) The DBIN method [10] (PSNR = 52.50 dB, SAM = 3.40, ERGAS = 1.09, SSIM = **0.994**); (g) The proposed **DHIF-Net** method (PSNR = **52.64** dB, SAM = **3.37**, ERGAS = **1.08**, SSIM = **0.994**).

TABLE VII
AVERAGE PSNR, SAM, ERGAS, AND SSIM RESULTS OF THE TEST METHODS FOR GAUSSIAN NOISE ON THE CAVE DATASET AND THE HARVARD DATASET

Factor	8									
Dataset	CAVE									
Noise level	σ=10					σ=30				
Method	NSSR [24]	MHF-net [31]	DBIN [10]	DHIF-Net (ours, $T=1$)	DHIF-Net (ours, $T=4$)	NSSR [24]	MHF-net [31]	DBIN [10]	DHIF-Net (ours, $T=1$)	DHIF-Net (ours, $T=4$)
PSNR	33.65	36.22	35.88	36.28	**37.12**	27.44	32.24	32.48	32.52	**33.44**
SAM	22.34	11.45	8.22	8.75	**7.63**	31.80	15.03	11.77	11.69	**10.47**
ERGAS	2.70	2.00	2.08	1.98	**1.80**	5.76	3.13	2.98	2.98	**2.69**
SSIM	0.890	0.924	0.944	0.950	**0.955**	0.565	0.848	0.905	0.909	**0.920**
Dataset	Harvard									
Noise level	σ=10					σ=30				
Method	NSSR [24]	MHF-net [31]	DBIN [10]	DHIF-Net (ours, $T=1$)	DHIF-Net (ours, $T=4$)	NSSR [24]	MHF-net [31]	DBIN [10]	DHIF-Net (ours, $T=1$)	DHIF-Net (ours, $T=4$)
PSNR	31.84	33.62	36.50	38.86	**39.79**	28.76	30.19	34.49	35.10	**36.33**
SAM	18.63	14.95	5.66	5.31	**4.86**	22.27	12.80	6.75	7.07	**5.87**
ERGAS	6.35	4.43	2.68	2.18	**1.99**	8.24	5.13	3.23	3.03	**2.74**
SSIM	0.691	0.797	0.912	0.948	**0.953**	0.566	0.671	0.889	0.901	**0.912**
Factor	16									
Dataset	CAVE									
Noise level	σ=10					σ=30				
Method	NSSR [24]	MHF-net [31]	DBIN [10]	DHIF-Net (ours, $T=1$)	DHIF-Net (ours, $T=4$)	NSSR [24]	MHF-net [31]	DBIN [10]	DHIF-Net (ours, $T=1$)	DHIF-Net (ours, $T=4$)
PSNR	32.05	34.48	34.99	35.07	**36.14**	27.00	31.14	31.38	32.24	**32.88**
SAM	25.19	13.02	10.01	9.97	**10.18**	33.28	17.19	13.00	12.48	**11.21**
ERGAS	1.54	1.24	1.13	1.11	**1.05**	2.97	1.76	1.67	1.52	**1.40**
SSIM	0.882	0.907	0.923	0.947	**0.947**	0.563	0.822	0.870	0.908	**0.918**
Dataset	Harvard									
Noise level	σ=10					σ=30				
Method	NSSR [24]	MHF-net [31]	DBIN [10]	DHIF-Net (ours, $T=1$)	DHIF-Net (ours, $T=4$)	NSSR [24]	MHF-net [31]	DBIN [10]	DHIF-Net (ours, $T=1$)	DHIF-Net (ours, $T=4$)
PSNR	31.50	33.20	36.14	38.30	**39.18**	28.49	29.94	34.65	34.95	**36.17**
SAM	20.14	10.68	5.86	5.78	**5.07**	22.97	11.19	6.82	7.57	**6.09**
ERGAS	3.56	2.17	1.36	1.17	**1.06**	4.16	2.54	1.59	1.57	**1.38**
SSIM	0.671	0.765	0.909	0.945	**0.950**	0.568	0.636	0.891	0.896	**0.912**

the best five competing methods in this comparison group. We can see that the proposed method can recover more details of the textures and edges than the other competing methods. The MHF-net method [31] and the DBIN method [10] used standard convolution to learn image priors including the spatio-spectral dependency, that is, all positions of the image share the same convolution kernel. However, in our method, we propose a deep neural network to estimate the spatially varying 3D filters. These spatially adaptive 3D filters are input-dependent and have advantages in capturing the rich dynamics of spatio-spectral dependency. Therefore, for scenes with rich or complex image contents or textures, our proposed method is more robust than the MHF-net method [31] and the DBIN method [10] as shown in Figs. 3 and 4.

We have also conducted more experiments for Gaussian noise on the CAVE dataset and the Harvard dataset. We add the Gaussian white noise to the LR-HSIs and the HR-MSIs with noise level σ ($\sigma = 10, 30$). Since the hyperparameters of most model-based methods are only applicable to the noise-free scenes, we only selected the NSSR method [24], the MHF-net method [31], and the DBIN method [10] as the competing methods. From Table VII, we can see that our proposed method with four stages has the best PSNR results compared with the NSSR method [24], the MHF-net method [31], the DBIN method [10],

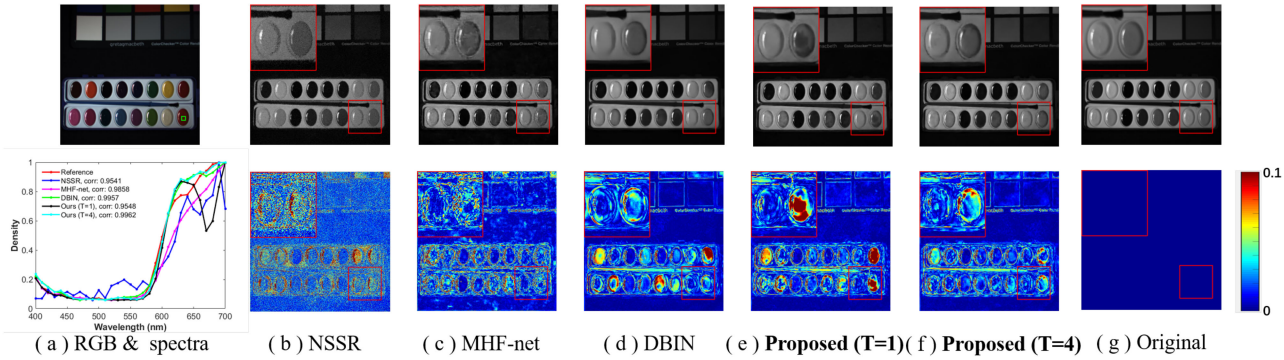


Fig. 7. Reconstructed images of $paints$ in CAVE dataset at 670 nm with noise level $\sigma = 30$ and scaling factor $s = 8$. The first row shows the reconstructed images, and the second row shows the error images of the competing methods. (a) The RGB image and the reconstructed spectra of the selected patch (indicated by a green box); (b) The NSSR method [24] (PSNR = 26.49 dB, SAM = 24.58, ERGAS = 3.24, SSIM = 0.613); (c) The MHF-net method [31] (PSNR = 30.66 dB, SAM = 10.90, ERGAS = 1.97, SSIM = 0.892); (d) The DBIN method [10] (PSNR = 30.20 dB, SAM = 8.79, ERGAS = 2.10, SSIM = 0.939); (e) The proposed **DHIF-Net** ($T = 1$) method (PSNR = 29.87 dB, SAM = 10.17, ERGAS = 2.18, SSIM = 0.930); (f) The proposed **DHIF-Net** ($T = 4$) method (PSNR = **31.20** dB, SAM = **8.19**, ERGAS = **1.87**, SSIM = **0.943**); (g) The ground truth images.

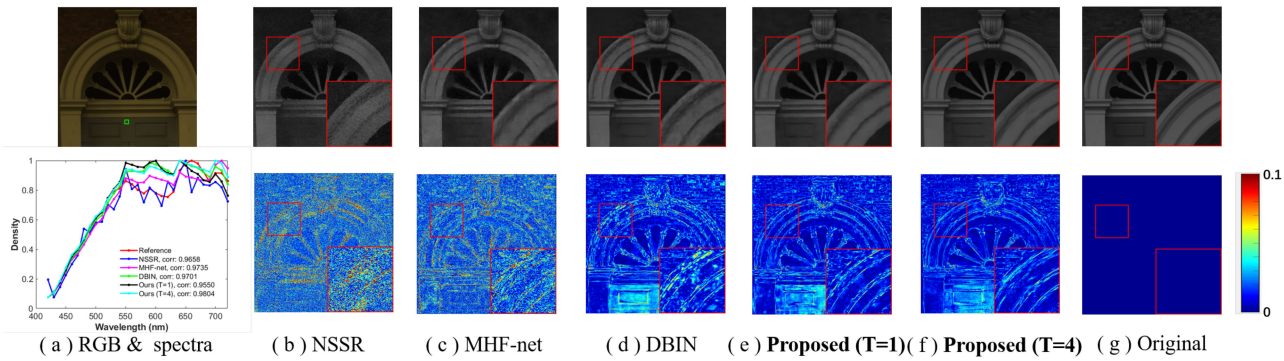


Fig. 8. Reconstructed images of $imge4$ in HARVARD dataset at 610 nm with noise level $\sigma = 30$ and scaling factor $s = 8$. The first row shows the reconstructed images, and the second row shows the error images of the competing methods. (a) The RGB image and the reconstructed spectra of the selected patch (indicated by a green box); (b) The NSSR method [24] (PSNR = 28.52 dB, SAM = 10.82, ERGAS = 4.85, SSIM = 0.517); (c) The MHF-net method [31] (PSNR = 31.28 dB, SAM = 4.88, ERGAS = 2.16, SSIM = 0.716); (d) The DBIN method [10] (PSNR = 36.82 dB, SAM = 3.12, ERGAS = 1.34, SSIM = 0.934); (e) The proposed **DHIF-Net** ($T = 1$) method (PSNR = 37.22 dB, SAM = 3.34, ERGAS = 1.17, SSIM = 0.940); (f) The proposed **DHIF-Net** ($T = 4$) method (PSNR = **38.41** dB, SAM = **2.79**, ERGAS = **1.11**, SSIM = **0.945**); (g) The ground truth images.

and the proposed method with one stage. On the CAVE dataset, the proposed method outperforms the MHF-net method [31] and the DBIN method [10] can be up to 1.66 dB and 1.15 dB for the Gaussian noise with noise level $\sigma = 10$ and scaling factor $s = 16$, respectively. On the Harvard dataset, the PSNR results gain over the MHF-net method [31] can be up to over 6 dB for the Gaussian noise with noise levels $\sigma = 10$ and 30, respectively. On the Harvard dataset, the proposed method outperforms the DBIN method [10] can be up to over 3 dB for the Gaussian noise with noise level $\sigma = 10$ and scaling factors $s = 8$ and 16, respectively. Using four stages in our proposed network has better results compared to one stage. On the Harvard dataset, the PSNR results of four stages gain over one stage can be up to 1.23 dB and 1.22 dB for the scaling factors 8 and 16 with noise level $\sigma = 30$, respectively. Figs. 7, 8, and 9 show the visual quality comparison results of the competing methods. From Figs. 7, 8, and 9, we can see that our proposed method removes the noise effectively and has less undesirable visual artifacts. We also show the colorchecker reconstructed by the competing methods in Fig. 10. It can be observed that our proposed method can suppress more noise and have more consistent colors with the ground truth image.

D. Experiment With Real Multispectral Images

We have also conducted experiments on a real MSI dataset, namely, WV2. The dataset contains a pair of a real LR-MSI of size $419 \times 658 \times 8$ and an HR-RGB image of size $1676 \times 2632 \times 3$. The upper half parts of the LR MSI and the HR-RGB image are used for training, and the remaining parts are used for testing. Since there is no real HR-MSI in WV2 dataset, we use the Wald's protocol [53] to generate the training samples. Under the Wald's protocol, the original LR-MSI and HR-RGB image are degraded spatially with a factor of 4 for generating the training inputs. The original LR-MSI is regarded as the reference. After training, the remaining parts of the LR-MSI and HR-RGB image are used to generate a HR-MSI by the test methods. We compare the proposed method with the HySure [47] method and the other two deep learning-based methods, i.e., the MHF-net method [31] and the DBIN method [10]. Fig. 11

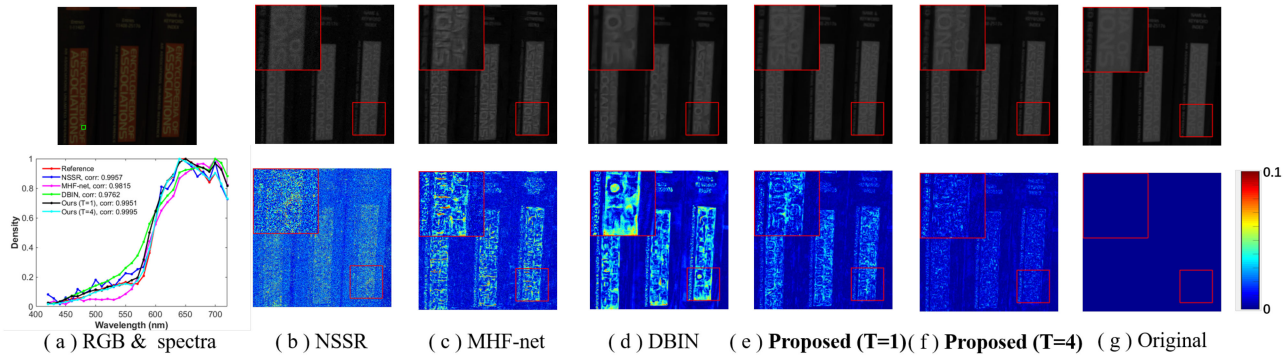


Fig. 9. Reconstructed images of $imgh2$ in HARVARD dataset at 620 nm with noise level $\sigma = 30$ and scaling factor $s = 16$. The first row shows the reconstructed images, and the second row shows the error images of the competing methods. (a) The RGB image and the reconstructed spectra of the selected patch (indicated by a green box); (b) The NSSR method [24] (PSNR = 30.62 dB, SAM = 35.46, ERGAS = 5.66, SSIM = 0.635); (c) The MHF-net method [31] (PSNR = 31.22 dB, SAM = 23.88, ERGAS = 4.58, SSIM = 0.656); (d) The DBIN method [10] (PSNR = 36.82 dB, SAM = 11.72, ERGAS = 2.62, SSIM = 0.924); (e) The proposed **DHIF-Net** ($T = 1$) method (PSNR = 37.02 dB, SAM = 12.94, ERGAS = 2.55, SSIM = 0.936); (f) The proposed **DHIF-Net** ($T = 4$) method (PSNR = **39.59** dB, SAM = **9.55**, ERGAS = **2.02**, SSIM = **0.956**); (g) The ground truth images.

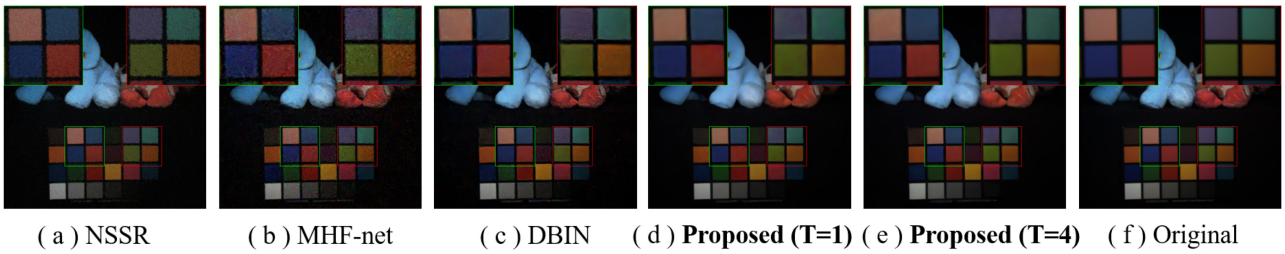


Fig. 10. Reconstructed HR RGB images are generated by applying the spectral response of a Nikon D700 camera. The reconstructed image by (a) The NSSR method [24]; (b) The MHF-net method [31]; (c) The DBIN method [10]; (d) The proposed **DHIF-Net** (T = 1) method; (e) The proposed **DHIF-Net** (T = 4) method; (f) The ground truth image;.

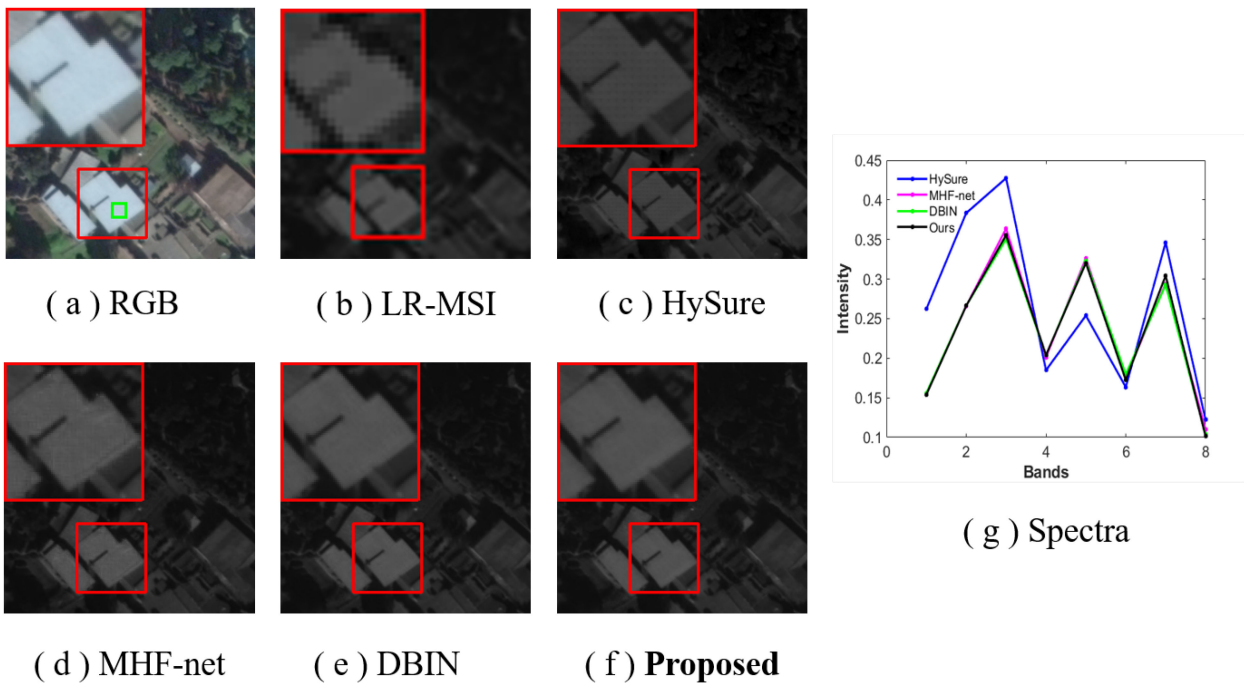


Fig. 11. Reconstructed images of WV2 of band 5. (a) The real HR RGB image; (b) The LR MSI; the reconstructed image by (c) The HySure method [47]; (d) The MHF-net method [31]; (e) The DBIN method [10]; (f) The proposed **DHIF-Net** method; (g) The reconstructed spectra of the selected patch (indicated by a green box).

TABLE VIII
AVERAGE PSNR RESULTS ON THE CAVE DATASET AND THE HARVARD DATASET FOR DISCUSSING THE MODEL GENERALIZATION

Training Set	CAVE			Harvard		
Testing Set	Harvard			CAVE		
Factor	8	16	32	8	16	32
Methods	PSNR	PSNR	PSNR	PSNR	PSNR	PSNR
DHSIS [49]	46.03	44.26	42.73	36.52	33.05	32.08
MHF-net [31]	43.82	42.87	41.98	34.32	32.07	30.47
DBIN [10]	45.20	42.21	39.43	32.57	31.04	29.37
DHIF-Net (ours)	**46.68**	**45.08**	**43.23**	**43.32**	**41.23**	**36.67**

TABLE IX
FLOPS, PARAMETERS AND PERFORMANCE OF THE COMPETING METHODS

Methods	TFLOPS	Params. (M)	PSNR	SAM	ERGAS	SSIM
MHF-net [31]	0.24	2.03	46.46	4.37	0.67	0.992
DBIN [10]	3.99	2.98	48.73	3.11	0.55	0.994
Proposed DHIF-Net-Small	2.69	3.11	49.43	3.10	0.52	0.995
Proposed DHIF-Net	3.62	22.67	49.79	3.01	0.51	0.995

shows parts of the reconstructed HR-MSIs. From Fig. 11, we can see that the proposed DHIF-Net method can better suppress undesirable visual artifacts and recover more details around edges and textures.

E. Model Generalization

To discuss the model generalization, we have conducted experiments when training on the CAVE dataset and testing on the Harvard dataset or training on the Harvard dataset and testing on the CAVE dataset, respectively. From Table VIII, we can see that our proposed method has the best generalization performance. The DHSIS method [49] is not a pure deep learning method. The DHSIS method trained a deep CNN to learn deep image priors firstly and inserted the learned deep image priors into the hyperspectral image fusion framework. Therefore, the DHSIS method is more robust than the DBIN method [10] and the MHF-net method [31]. By contrast, our method proposes a model-guided unfolding network to learn the rich dynamics of spatio-spectral dependency. The joint exploitation of spatio-spectral regularization and physical imaging models makes our method more robust and flexible than other competing ones in the open literature.

F. Complexity Analysis

In addition to the visual quality comparison, we have also compared with other deep learning-based methods in terms of computational complexity (measured by TFLOPS and number of network parameters) on the CAVE dataset. We have also implemented a variant of the proposed method (denoted as DHIF-Net-Small), where we set the number of feature channels to 64 for all convolutional layers in the U-net for 3D filter estimation. From Table IX, we can see that the proposed DHIF-Net-Small method has less TFLOPS and comparable number of parameters with the DBIN method, but still outperforms the DBIN method. When compared with the proposed DHIF-Net-Small, the proposed DHIF-Net method with a larger number of feature channels in the U-net has achieved the PSNR gain of 0.36 dB. This also indicates that there is great parameter redundancy in the backbone network for 3D filter prediction, which can be significantly reduced without significant performance loss. Recently developed neural architecture search (NAS) [54], [55] seems an appropriate framework for achieving an improved trade-off between cost and performance. We will leave it as future work.

VI. CONCLUSION

In this paper, we have proposed a new network named DHIF-Net based on iterative spatio-spectral regularization for HSI fusion. Through combining deep spatio-spectral regularization with a physical imaging model, we have developed an optimization-inspired algorithm for reconstructing the HR-HSI from a pair of HR-MSI and LR-HSI. We have unfolded this algorithm into a deep convolutional network implementation to adaptively estimate the spatially varying 3D filters for constructing the spatio-spectral regularization. Unlike previous works, spatially varying 3D filters can fully encode the spatial and spectral dependencies within the HSIs. Since the proposed optimization problem is differentiable, we can solve it in an end-to-end training manner. Experimental results on both synthetic datasets and real-world MSI data show that the proposed DHIF-Net outperforms other state-of-the-art competing methods. We will report additional experimental results on other modified hardware (e.g., [56], [57], and [58]) in the future.

REFERENCES

[1] M. Uzair, A. Mahmood, and A. S. Mian, "Hyperspectral face recognition using 3D-DCT and partial least squares." in *Proc. Brit. Mach. Vis. Conf.*, 2013, pp. 57.1–57.10.
[2] W. Xie, T. Jiang, Y. Li, X. Jia, and J. Lei, "Structure tensor and guided filtering-based algorithm for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4218–4230, 2019.

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

[4] B. Uzkent, M. J. Hoffman, and A. Vodacek, "Real-time vehicle tracking in aerial video using hyperspectral features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 36–44.

[5] B. Uzkent, A. Rangnekar, and M. Hoffman, "Aerial vehicle tracking by adaptive fusion of hyperspectral likelihood maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 39–48.

[6] H. Van Nguyen, A. Banerjee, and R. Chellappa, "Tracking via object reflectance using a hyperspectral video camera," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops*, 2010, pp. 44–51.

[7] 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.

[8] W. Xie, J. Lei, Y. Cui, Y. Li, and Q. Du, "Hyperspectral pansharpening with deep priors," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1529–1543, May 2020.

[9] W. Xie, Y. Cui, Y. Li, J. Lei, Q. Du, and J. Li, "HPGAN: Hyperspectral pansharpening using 3-D generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 463–477, Jan. 2021.

[10] W. Wang, W. Zeng, Y. Huang, X. Ding, and J. Paisley, "Deep blind hyperspectral image fusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 4150–4159.

[11] Q. Xie, M. Zhou, Q. Zhao, Z. Xu, and D. Meng, "MHF-Net: An interpretable deep network for multispectral and hyperspectral image fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1457–1473, Mar. 2022.

[12] V. K. Shettigara, "A generalized component substitution technique for spatial enhancement of multispectral images using a higher resolution data set," *Photogrammetric Eng. remote Sens.*, vol. 58, no. 5, pp. 561–567, 1992.

[13] B. Aiazzi, S. Baronti, F. Lotti, and M. Selva, "A comparison between global and context-adaptive pansharpening of multispectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 2, pp. 302–306, Apr. 2009.

[14] J. Nunez, X. Otazu, O. Fors, A. Prades, V. Pala, and R. Arbiol, "Multiresolution-based image fusion with additive wavelet decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 3, pp. 1204–1211, May 1999.

[15] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph.*, vol. 26, no. 3, pp. 96.1–96.5, 2007.

[16] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 556–562.

[17] D. D. Lee, "Learning the parts of objects by non-negative matrix factorization." *Nature*, vol. 401, pp. 788–791, 1999.

[18] 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.

[19] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "A new pansharpening algorithm based on total variation," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 318–322, Jan. 2014.

[20] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, "An iterative regularization method for total variation-based image restoration," *Multiscale Model. Simul.*, vol. 4, no. 2, pp. 460–489, 2005.

[21] N. Akhtar, F. Shafait, and A. Mian, "Sparse spatio-spectral representation for hyperspectral image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 63–78.

[22] C. Grohnfeldt, X. X. Zhu, and R. Bamler, "Jointly sparse fusion of hyperspectral and multispectral imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2013, pp. 4090–4093.

[23] Y. Zhao, J. Yang, Q. Zhang, L. Song, Y. Cheng, and Q. Pan, "Hyperspectral imagery super-resolution by sparse representation and spectral regularization," *EURASIP J. Adv. Signal Process.*, vol. 2011, no. 1, pp. 1–10, 2011.

[24] 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.

[25] L. Zhang, W. Wei, C. Bai, Y. Gao, and Y. Zhang, "Exploiting clustering manifold structure for hyperspectral imagery super-resolution," *IEEE Trans. Image Process.*, vol. 27, no. 12, pp. 5969–5982, Dec. 2018.

[26] C. Dong, C. C. Loy, K. He, and X. Tang, "Learn. a deep convolutional Netw. image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2014, pp. 184–199.

[27] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[28] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep CNN denoiser prior for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3929–3938.

[29] W. Dong, P. Wang, W. Yin, G. Shi, F. Wu, and X. Lu, "Denoising prior driven deep neural network for image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2305–2318, Oct. 2019.

[30] 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.

[31] 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. Vis. Pattern Recognit.*, 2019, pp. 1585–1594.

[32] X. Zhang, W. Huang, Q. Wang, and X. Li, "SSR-NET: Spatial-spectral reconstruction network for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5953–5965, Jul. 2021.

[33] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, Jun. 2013.

[34] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2864–2875, Jul. 2013.

[35] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 60–65.

[36] P. Coupé, P. Yger, S. Prima, P. Hellier, C. Kervrann, and C. Barillot, "An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images," *IEEE Trans. Med. Imag.*, vol. 27, no. 4, pp. 425–441, Apr. 2008.

[37] B. Mildenhall, J. T. Barron, J. Chen, D. Sharlet, R. Ng, and R. Carroll, "Burst denoising with kernel prediction networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2502–2510.

[38] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive separable convolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 261–270.

[39] K. Zhang, L. V. Gool, and R. Timofte, "Deep unfolding network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3217–3226.

[40] C. Bertocchi, E. Chouzenoux, M.-C. Corbineau, J.-C. Pesquet, and M. Prato, "Deep unfolding of a proximal interior point method for image restoration," *Inverse Problems*, vol. 36, no. 3, 2020, Art. no. 034005.

[41] Y. Chen and T. Pock, "Trainable Nonlinear Reaction Diffusion: A. Flexible Framework for Fast and Effective Image Restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1256–1272, Jun. 2017.

[42] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, "Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2241–2253, Sep. 2010.

[43] A. Chakrabarti and T. Zickler, "Statistics of Real-World Hyperspectral Images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 193–200.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Int. Conf. Learn. Representations*, 2015, pp. 1–13.

[45] 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. Vis.*, 2015, pp. 1026–1034.

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

[47] 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.

[48] 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.

[49] 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.

[50] Y. Zhang, S. De Backer, and P. Scheunders, "Noise-resistant wavelet-based Bayesian fusion of multispectral and hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 11, pp. 3834–3843, Nov. 2009.

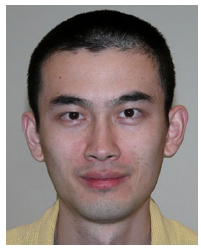
[51] L. Wald, "Quality of high resolution synthesised images: Is there a simple criterion?," in *Proc. 3rd Conf. Fusion Earth Data, Merging Point Meas., Raster Maps Remotely Sensed Images*, 2000, pp. 99-103.
[52] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[53] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of Satellite Images of Different Spatial Resolutions: Assessing the Quality of Resulting Images" *Photogrammetric Eng. Remote Sens.*, vol. 63, no. 6, pp. 691–699, 1997.
[54] T. Elsken *et al.*, "Neural architecture search: A survey." *J. Mach. Learn. Res.*, vol. 20, no. 55, pp. 1–21, 2019.
[55] A. Wan *et al.*, "FBNetV2: Differentiable neural architecture search for spatial and channel dimensions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12965–12974.
[56] X. Cao, X. Tong, Q. Dai, and S. Lin, "High resolution multispectral video capture with a hybrid camera system," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 297–304.
[57] C. Ma, X. Cao, R. Wu, and Q. Dai, "Content-adaptive high-resolution hyperspectral video acquisition with a hybrid camera system," *Opt. Lett.*, vol. 39, no. 4, pp. 937–940, 2014.
[58] V. Saragadam, M. Dezeeuw, R. G. Baraniuk, A. N. Veeraraghavan, and A. Sankaranarayanan, "SASSI-super-pixelated adaptive spatio-spectral imaging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 7, pp. 2233–2244, Jul. 2021.

Tao Huang received the B.S. degree in electronic engineering in 2018 from Xidian University, Xi'an, China, where he is currently working toward the Ph.D. degree in electronic science and technology. His research interests include image restoration and deep learning.

Weisheng Dong (Member, IEEE) received the B.S. degree in electronic engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2004 and the Ph.D. degree in circuits and system from Xidian University, Xi'an, China, in 2010. In 2006, he was a Visiting Student with Microsoft Research Asia, Beijing, China. From 2009 to 2010, he was a Research Assistant with the Department of Computing, Hong Kong Polytechnic University, Hong Kong. In 2010, he joined the School of Electronic Engineering, Xidian University, as a Lecturer, where he has been a Professor since 2016. His research interests include inverse problems in image processing, sparse signal representation, and image compression. He was the recipient of the Best Paper Award at the SPIE Visual Communication and Image Processing in 2010. He is currently an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING.

Jinjian Wu (Member, IEEE) received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2008 and 2013, respectively. From 2011 to 2013, he was a Research Assistant with Nanyang Technological University, Singapore, where he was a Postdoctoral Research Fellow from 2013 to 2014. From 2015 to 2019, he was an Associate Professor with Xidian University, where he has been a Professor since 2019. His research interests include visual perceptual modeling, biomimetic imaging, quality evaluation, and object detection. He was the recipient of the Best Student Paper Award at ISCAS in 2013. He was an Associate Editor for the *Journal of Circuits, Systems, and Signal Processing*, the Special Section Chair of IEEE Visual Communications and Image Processing in 2017, and the Section Chair, an Organizer, and the TPC Member for ICME from 2014 to 2015, PCM from 2015 to 2016, ICIP in 2015, VCIP in 2018, and AAAI in 2019.

Leida Li (Member, IEEE) received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2004 and 2009, respectively. In 2008, he was a Research Assistant with the Department of Electronic Engineering, Kaohsiung University of Science and Technology, Kaohsiung, Taiwan. From 2014 to 2015, he was a Visiting Research Fellow with the Rapid-Rich Object Search Lab, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, where he was a Senior Research Fellow from 2016 to 2017. He is currently a Professor with the School of Artificial Intelligence, Xidian University. His research interests include multimedia quality assessment, affective computing, information hiding, and image forensics. From 2019 to 2021, he was the SPC of IJCAI, the Session Chair for ICMR in 2019 and PCM in 2015, and the TPC Member for CVPR in 2021, ICCV in 2021, AAAI from 2019 to 2021, ACM MM from 2019 to 2020, ACM MM-Asia in 2019, ACII in 2019, and PCM in 2016. He is also an Associate Editor for the *Journal of Visual Communication and Image Representation* and *EURASIP Journal on Image and Video Processing*.



Xin Li (Fellow, IEEE) received the B.S. (Hons.) degree in electronic engineering and information science from the University of Science and Technology of China, Hefei, China, in 1996, and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, USA, in 2000. From 2000 to 2002, he was a Technical Staff Member with the Sharp Laboratories of America, Camas, WA, USA. Since 2003, he has been a Faculty Member with the Lane Department of Computer Science and Electrical Engineering. His research interests include image/video coding and processing. He was the recipient of the Best Student Paper Award at the Conference of Visual Communications and Image Processing in 2001, Best Student Paper Award at the IEEE Asilomar Conference on Signals, Systems and Computers in 2006, and Best Paper Award at the Conference of Visual Communications and Image Processing in 2010. He is currently a Member of the Image, Video, and Multidimensional Signal Processing Technical Committee and an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.

Guangming Shi (Fellow, IEEE) received the B.S. degree in automatic control, the M.S. degree in computer control, and the Ph.D. degree in electronic information technology from Xidian University, Xi'an, China, in 1985, 1988, and 2002, respectively. In 1988, he joined the School of Electronic Engineering, Xidian University. From 1994 to 1996, as a Research Assistant, he cooperated with the Department of Electronic Engineering, The University of Hong Kong, Hong Kong. Since 2003, he has been a Professor with the School of Electronic Engineering, Xidian University. In 2004, he was the Head of National Instruction Base of Electrician and Electronic. From June to December 2004, he studied with the Department of Electronic Engineering, University of Illinois at Urbana-Champaign, Champaign, IL, USA. He is currently the Vice President with Xidian University, and the Academic Leader in the subject of circuits and systems. He has authored or coauthored more than 60 research papers. His research interests include compressed sensing, theory and design of multirate filter banks, image denoising, low-bit-rate image/video coding, and implementation of algorithms for intelligent signal processing (using DSP and FPGA).