Hyperspectral Snapshot Compressive Imaging With Dense Back-Projection Joint Attention Network

Yubao Sun, Junru Huang, Liling Zhao, and Kai Hu

Abstract—**The hyperspectral snapshot compressive imaging (SCI) system encodes three-dimensional hyperspectral images into a single two-dimensional snapshot measurement and then decodes the underlying 3-D hyperspectral images by solving the compressive sensing reconstruction problem. Practical applications of SCI imaging systems require fast and high-quality reconstruction. To meet this requirement, we propose a novel encoder and decoder network with dense back-projection joint attention for hyperspectral SCI. The main contributions of our network lie in two aspects. First, we propose a dense back-projection module and deploy it in an encoder with five scales. It computes the back-projection between each scale and all its preceding scales, thereby fusing complementary information between different scales for efficient reconstruction. Second, we design a spatial-spectral attention module and deploy it in the decoder to boost reconstruction quality. By exploiting a cascade of spatial-spectral attention, it can efficiently capture spatial and spectral correlations in hyperspectral images with a low volume of parameters. In addition, a compound loss, including the reconstruction loss, and the spatial–spectral total variation loss, is designed to guide network learning in an end-to-end manner. Intensive experiments on simulation and real data show that our method has obvious advantages over multiple state-of-the-art methods, achieving a significant improvement in reconstruction quality and a substantial reduction in running time.**

Index Terms—**Compound loss, compressive spectral imaging, dense back-projection (DBP), lightweight network, spatial–spectral attention (SSA).**

I. Introduction

HYPERSPECTRAL image (HSI) is a three-dimensional (3-D) data cube, which contains a number of 2-D spectral bands. The spectral bands in HSI are obtained by dense sampling at small intervals within a certain wavelength range. Therefore, each pixel represents a spectral signature, which can be used to distinguish different types of objects in the scene. HSI has wide application in land object classification, scientific experiments, industry detection, and other fields [1]–[4].

Hyperspectral imaging is the first step before HSI-based applications. HSIs usually use 1- or 2-D detectors to collect the reflectance of the scene with respect to different wavelengths based on Shannon sampling method. In order to cover all three dimensions of HSI images, spatial or spectral scanning is required over multiple exposure periods [5]. However, the scanning operation increases the imaging time while requiring the scene to be static during imaging to avoid inconsistencies in scene content [6]. Unlike these Shannon-based acquisition methods, snapshot compressive imaging (SCI) systems based on compressed sensing theory [7] have been proposed to capture dynamic objects and scenes. Among these systems, coded aperture snapshot spectral imaging (CASSI) is a promising solution. By introducing coded aperture [8], CASSI first encodes 3-D HSI into a 2-D snapshot measurement in a single exposure without scanning operations. Then, the underlying HSI is reconstructed from the captured 2-D snapshot measurements by solving an optimization problem, termed hyperspectral snapshot compressive reconstruction. Due to the underdetermined sampling mechanism of the CASSI system, hyperspectral snapshot compressive reconstruction is a highly ill-posed problem. In order to solve this problem, earlier works modeled hyperspectral snapshot compressive reconstruction as a prior regularized optimization problem. Some priors, such as total variation (TV) [9], [10], sparse representation [11], and nonlocal self-similarity [12] have been designed to represent HSIs. However, these priors are generally manually designed based on simplified assumptions and cannot effectively represent complex spatial–spectral structures in HSIs. In addition, these prior regularization driven methods require iterative optimization and are computationally complex. Recently, deep networks [13] have been widely recognized for outstanding feature representation capabilities. Many works turned to deep learning driven methods and enabled hyperspectral snapshot compressive reconstruction by learning an inverse mapping from snapshot measurements to original HSIs. Some representative methods include ISTA-Net [14], SSR-Net [15], NSSR-Net [16], and so on [17]. Although these deep learning-based methods can directly output reconstructions through one-shot feed-forward network computations, their reconstruction quality still needs to be further improved. Some other works are devoted to developing more complex deep networks to improve reconstruction quality, such as using nonlocal attention [16] and 3-D convolution. However, employing complex networks increases the reconstruction time inevitably.

To cope with the abovementioned issue, this work aims to investigate a lightweight network capable of fast and high-quality reconstruction. Specifically, we propose a novel dense back-projection (DBP) joint attention network (dubbed as DBPA-Net) to learn the parametric reconstruction mapping. First, we design

Manuscript received 30 April 2022; revised 30 June 2022; accepted 6 July 2022. Date of publication 19 July 2022; date of current version 4 August 2022. This work was supported by the National Natural Science Foundation of China under Grant U2001211. *(Corresponding author: Kai Hu.)*

The authors are with the Engineering Research Center of Digital Forensics, Ministry of Education, and the Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: sunyb@nuist.edu.cn; 20201249083@nuist.edu.cn; zhaoliling@nuist.edu.cn; 001600@nuist.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3192484

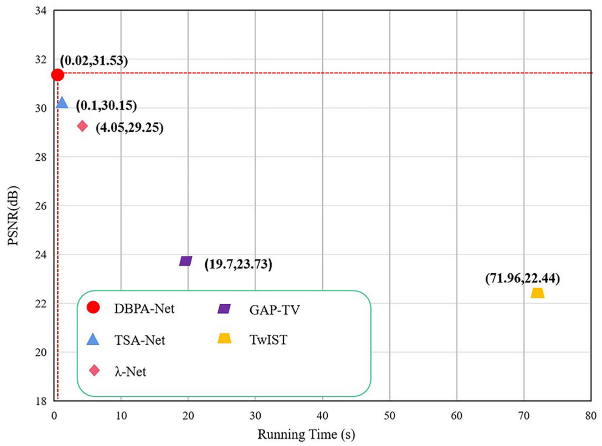


Fig. 1. Running time and PSNR metrics comparison between the proposed DBPA-Net and multiple state-of-the-art methods.

a DBP module to perform DBP operations between each scale and all its preceding scales. The DBP between the current scale and its preceding scales can fuse the high-resolution feature maps into the current scale, thereby compensating for the loss of information in the down-sampling operation. The fused features are then transmitted to the decoder for improving reconstruction quality. Second, we design a spatial–spectral attention (SSA) module to capture the spatial–spectral correlations in HSIs. This attention module is lightweight and can bring reconstruction performance improvement. As shown in Fig. 1, compared with the state-of-the-art methods, the proposed DBPA-Net has the best reconstruction quality with the lowest running time. Our main contributions can be summarized as follows.

1) The proposed DBP module links each scale and all its preceding scales, and fuses complementary features between different scales through back-projection. Therefore, it can effectively compensate for the lost high-resolution spatial feature information at each scale to improve reconstruction performance.
2) We design an attention module to capture the spatial–spectral correlation of HSI for reconstruction. This module is lightweight and effective by employing the cascade of spatial and spectral attention.
3) We further exploit the efficiency of these two modules for reconstruction in a multiscale manner by deploying them into an encoder and decoder architecture. The experimental results verify that DBPA-Net has made significant progress in both reconstruction time and reconstruction quality.

The rest of this article is organized as follows. Section II reviews the related works. Section III describes the forward model of single disperser CASSI. Section IV presents the proposed method. The experimental results are provided in Section V. Finally, Section VI concludes this article.

II. Related Work

Hyperspectral SCI systems encode the spatial and spectral information as 2-D snapshot measurements and the desired 3-D HSI can only be obtained by conducting a reconstruction algorithm. Many algorithms have been proposed to solve this problem, mainly including prior-driven and network-driven two categories.

A. Prior-Driven Reconstruction

Due to the inherent underdetermined sampling, the reconstruction algorithm is an ill-posed inverse problem. Prior-driven reconstruction methods mainly exploit different HSI priors to regularize this inverse problem. Typically, this category of methods formulates the reconstruction problem as a convex optimization with a prior regularization and fidelity term, and then the optimal solution is found by iterative optimization. Therefore, designing an appropriate prior plays a key role in prior-driven methods.

Sparse prior is widely used for hyperspectral SCI reconstruction, and the gradient projection sparse reconstruction (GPSR) [18] is a representative method using sparse prior. Specifically, GPSR imposed sparse constraints on the whole 3-D HSI. By constraining the sparsity in the image gradient domain, TV [9] prior was used to eliminate noise in the reconstruction image. TwIST [19] also used TV prior as the regularization term and realized compressive sensing reconstruction by a two-step iterative shrinkage thresholding algorithm. GAP-TV [20] used the generalized alternating projection (GAP) algorithm to optimize the reconstruction of HSIs. TVAL3 further solved the TV regularized least squares problem alternating direction method of multipliers [21], [22]. Liu et al. [23] proposed the DeSCI method, which uses the weighted nuclear norm to characterize the low rank prior of a group of matched patches. At the same time, an alternating minimization algorithm was developed to solve such problems.

However, this category of reconstruction methods needs many iterations to solve the inverse problem, which brings high time complexity [24]. In addition, the image prior and regularization parameters involved in the algorithm need to be carefully set manually, and the reconstruction quality needs to be improved.

B. Network-Driven Reconstruction

The network-driven methods use the powerful learning ability of the deep network to realize reconstruction [25]. This category of methods is mainly based on supervised learning and learns the explicit reconstruction mapping directly from the snapshot measurement to HSI by training the network on a large number of training samples. Different from the iterative optimization method, the network-driven method can reconstruct HSI by only performing a feedforward calculation on the learned network. Therefore, the key to improving the reconstruction performance of such methods is how to design an effective network.

Here, we introduce some representative HSI reconstruction networks. Xiong et al. [26] designed a convolution neural network to learn the hyperspectral compressive reconstruction. This work first up-sampled the measurement to make it have the same dimension as the original HSI, and then the reconstruction was enhanced by using a convolution neural network to learn incremental residuals. Choi et al. designed an autoencoder to obtain the nonlinear spectral representation of HSIs and used it as a spectral prior to the variational models [27]. Miao et al. [28]

proposed a λ-Net network to learn the compressive sensing reconstruction of HSIs and videos generated in two stages against the network. Wang et al. [29] designed a HyperRecon-Net to learn the reconstruction of HSIs, cascading spatial networks and inter spectral networks to predict the spectral information in HSIs. In the first stage of reconstruction, U-Net including a self-attention mechanism was used, and in the second stage, another U-Net was used to improve the reconstruction quality. Meng et al. [30] proposed TSA-Net and captured spatial–spectral correlation by independent similarity computation inside each dimension. TSA-Net had a reasonable computation cost. In general, the core requirement of network-driven reconstruction is to achieve both high-quality and low-cost reconstruction. Although these deep learning methods can reduce the reconstruction time, the reconstruction quality still needs to be improved.

C. Attention Mechanism

In order to better capture the correlation between spatial pixels and spectral bands in HSI, the network-driven methods began to introduce attention mechanisms to various tasks of HSIs [31]–[33]. Hu et al. proposed squeeze-and-excitation (SE) attention to adjust the weights of different feature map channels by learning the relationship between channels, therefore improving the expressive ability of the network [34]. Woo et al. [35] proposed the convolutional block attention module (CBAM) to concatenate channel attention and spatial attention. Zhang *et al.* proposed an efficient shuffle attention (SA) module, which uses shuffle units to effectively combine the two attention mechanisms. For hyperspectral snapshot compressive reconstruction [36], Meng et al. [30] used the spatial–spectral self-attention (TSA) to process the feature information from the channel dimension and the spatial dimension, respectively. Yang *et al.* [16] employed a nonlocal spatial attention module to capture the long-range dependencies in space, achieving high-quality reconstruction. In recent years, the attention mechanism has been proved to offer great potential in improving the performance of deep convolutional neural networks (CNNs). However, most existing methods focus on developing more complex attention modules for a better performance. This increases the complexity of the network model inevitably.

The proposed SSA module in this article is also related to the attention mechanism, but it is designed for lightweight CNNs. We design a joint network with a spatial residual attention block and a lightweight spectral attention block to extract spectral-spatial features. This attention mechanism is used to emphasize meaningful features along with the two blocks. Our proposed SSA module only involves a handful of parameters while bringing clear performance gain.

III. CASSI Forward Model

The CASSI systems mainly contain two categories. The first category encodes in the spatial domain, such as SD-CASSI [37]. The second category encodes in both spatial and spectral domains, including SS-CASSI [38], DD-CASSI [39]. Before detailing the proposed reconstruction network, we first briefly introduce the forward model of the SD-CASSI system, that is, the mathematical formula that describes the acquisition of snapshot measurements.

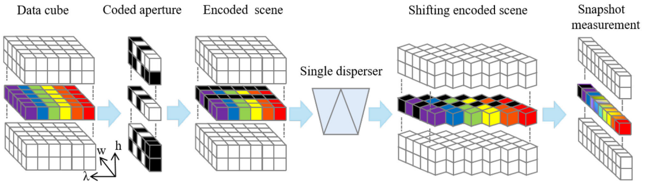


Fig. 2. Diagram of SD-CASSI system. It sequentially encodes, shifts, and integrates the 3-D incident scene, and finally obtains the 2-D snapshot measurements.

Fig. 2 shows the diagram of the SD-CASSI system. The scene is projected onto the coded aperture as spatial modulation. Then, the encoded scene passes through the prism and shifts the spectral bands. Last, the shifted version of the encoded scene is integrated along the spectral dimension by the detector, which results in the 2-D snapshot measurements. Let $\mathbf{X}(h, w, \lambda)$ indicates the 3-D spatial–spectral cube corresponding to the incoming scene, where $1 \leq h \leq H$ and $1 \leq w \leq W$ denote the spatial dimensions and $1 \leq \lambda \leq \Lambda$ represents the spectral dimension. The spatial modulation operation determined by coded aperture is described as the transmission function $F(h, w)$, and the dispersion function of the prism $\psi(\lambda)$ takes wavelength as its parameter. Mathematically, the captured 2-D snapshot measurements $\mathbf{Y} \in R^{H\times(W+\Lambda-1)}$ is formulated as

$$\mathbf{Y}(h,w) = \sum_{\lambda=1}^{\Lambda} F\left(h, w+\psi\left(\lambda\right)\right) \odot \mathbf{X}\left(h, w+\psi\left(\lambda\right), \lambda\right) \quad (1)$$

where $\odot$ represents the element-wise multiplication. In (1), the shifting is along the w-axis.

Let vec$(\cdot)$ denotes the operation of concatenating all the columns of a matrix as one single vector. Denote $y = \text{vec}(\mathbf{Y})$, $\hat{\mathbf{x}}_\lambda = \text{vec}(\mathbf{X}(:,:,\lambda))$, and $\mathbf{x} = \text{vec}((\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_\Lambda))$. According to the CASSI imaging principle, the snapshot measurement matrix $\boldsymbol{\Psi} \in R^{H(W+\Lambda-1)\times HW\Lambda}$ is determined by the coded aperture pattern $F(h, w)$, and the captured snapshot measurement can be written as

$$\mathbf{y} = \boldsymbol{\Psi}\mathbf{x} + \varepsilon \quad (2)$$

where $\mathbf{y} \in R^{H(W+\Lambda-1)}$ and $\mathbf{x} \in R^{HW\Lambda}$ are the vectorized representation of the snapshot measurement $\mathbf{Y}$ and the original HSI $\mathbf{X}$, and ε is noise.

The dimension of $\mathbf{y}$ is usually much lower than the dimension of $\mathbf{x}$. The obtained 2-D snapshot measurements are used to reconstruct the original HSIs by a reconstruction algorithm based on compressive sensing theory.

IV. DBP Joint Attention Network

In this section, we present a detailed illustration of the proposed network, which consists of a DBP encoder and SSA boosted decoder. As shown in Fig. 3, the proposed network takes the snapshot measurement as input and reconstructs the underlying HSI. It can be regarded as a function mapping from the snapshot measurement to the reconstruction. Encoder–decoder architecture is commonly used for representation learning, and

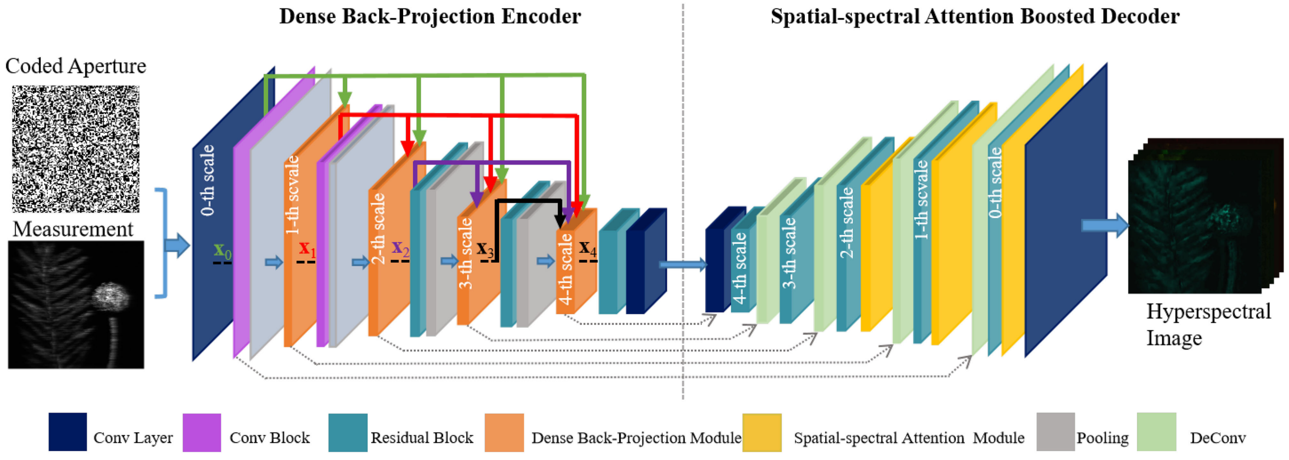


Fig. 3. Overall framework of the proposed DBPA-Net. DBPA-Net takes the coded aperture and snapshot measurement as the input and output the reconstruction. Four DBP modules are introduced into the encoder stage to enrich the information in each scale. Three SSA modules are introduced into the decoder to make full use of spatial–spectral correlation.

TABLE I
Configuration Details of the Proposed DBPA-Net

Output size	DBP encoder	Output size	Spatial-spectral Attention Boosted Decoder
$256\times256\times32$	$\text{conv}, 1\times1, 32, \text{stride}=1$	$256\times256\times28$	$\left[\begin{array}{c} conv, 1\times1, 28, stride=1 \\ Sigmoid\,(\cdot) \end{array}\right]$
$128\times128\times64$	$\left[\begin{array}{c} \text{conv}, 3\times3, 64, \text{stride}=1 \\ BN, ReLU \end{array}\right]*2$ $\text{maxpool}\,2d, 2\times2, \text{stride}=2$	$256\times256\times64$	$H_{\text{Spe}}\left(H_{\text{Spa}}\left(\cdot\right)\right), [64, 64]$ $\left[\begin{array}{c} Cat, Res2Net \\ conv, 3\times3, 64, stride=1 \\ BN, ReLU \end{array}\right]$ $convtranspose2d, 2\times2, stride=2$
$64\times64\times128$	$F_n, [64, 64]$ $\left[\begin{array}{c} \text{conv}, 3\times3, 128, \text{stride}=1 \\ BN, ReLU \end{array}\right]*2$ $\text{maxpool}\,2d, 2\times2, \text{ stride}=2$	$128\times128\times128$	$H_{\text{Spe}}\left(H_{\text{Spa}}\left(\cdot\right)\right), [128, 128]$ $\left[\begin{array}{c} Cat, Res2Net \\ conv, 3\times3, 128, stride=1 \\ BN, ReLU \end{array}\right]$ $convtranspose2d, 2\times2, stride=2$
$32\times32\times256$	$F_n, [128, 128]$ $\left[\begin{array}{c} \text{conv}, 3\times3, 256, \text{stride}=1 \\ BN, ReLU \\ Res2Net \end{array}\right]$ $\text{maxpool}\,2d, 2\times2, \text{stride}=2$	$64\times64\times256$	$H_{\text{Spe}}\left(H_{\text{Spa}}\left(\cdot\right)\right), [256, 256]$ $\left[\begin{array}{c} Cat, Res2Net \\ conv, 3\times3, 256, stride=1 \\ BN, ReLU \end{array}\right]$ $convtranspose2d, 2\times2, stride=2$
$16\times16\times512$	$F_n, [256, 256]$ $\left[\begin{array}{c} \text{conv}, 3\times3, 512, \text{stride}=1 \\ BN, ReLU \\ Res2Net \end{array}\right]$ $\text{maxpool}\,2d, 2\times2, \text{stride}=2$	$32\times32\times512$	$\left[\begin{array}{c} Cat, Res2Net \\ conv, 3\times3, 512, stride=1 \\ BN, ReLU \end{array}\right]$ $convtranspose2d, 2\times2, stride=2$
$16\times16\times1024$	$F_n, [512, 512]$ $\left[\begin{array}{c} \text{conv}, 3\times3, 1024, \text{stride}=1 \\ BN, ReLU \\ Res2Net \end{array}\right]$ $\text{conv}, 3\times3, 1024, \text{stride}=1$	$16\times16\times1024$	$\left[\begin{array}{c} Cat, Res2Net \\ conv, 3\times3, 1024, stride=1 \\ BN, ReLU \end{array}\right]$ $conv, 3\times3, 1024, stride=1$

F_n denotes DBP module, $H_{\text{Spe}}(\cdot)$ and $H_{\text{Spa}}(\cdot)$ denote attention module.

the encoder usually adopts the down-sampling operation to extract multiscale features. However, the down-sampling operation will cause the loss of detailed information, which is not conducive to the reconstruction task. To tackle this issue, the DBP encoder sets up dense connections between the current scale and its all preceding scales and fuses the complementary information of preceding to enhance the current scale by back-projection operation [40]. These fused features are then transmitted to the decoder for high-quality reconstruction. Furthermore, we leverage a lightweight self-attention module to jointly capture spatial-spectral correlations in HSIs and enhance the decoder to reconstruct more details. Table I lists the configuration details of the proposed network.

A. DBP Encoder

The DBP encoder takes the snapshot measurement as input and has five scales to extract hierarchical features. In the first scale, a 1×1 convolution is used to process the input snapshot measurement and generates a feature map $\mathbf{x_0}$ with 32 channels. $\mathbf{x_0}$ is the initial reconstruction, which can be regarded as the 0th scale. Then, $\mathbf{x_0}$ is processed by a conv block and a pooling

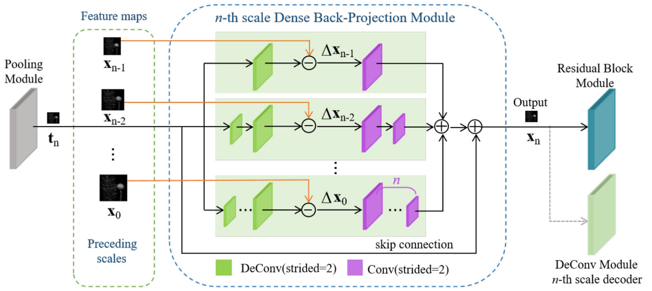


Fig. 4. Diagram of the nth scale DBP module. It contains n pairs of up-sampling and down-sampling operations to fuse complementary features (green box in the figure).

operation [41]. The two to five scales are configured with a DBP module, a convolution block, a residual block, a pooling operation, and a conv layer. Fig. 4 shows the internal diagram of the nth scale DBP module. The main function of this module is to fuse the complementary information between the nth scale and all its preceding scales and transmit the fused feature to the nth scale of the decoder to enrich reconstruction details. The function mapping F_{n} of this module is defined as

$$\mathbf{x_n} = F_{\mathrm{n}}\left(\mathbf{t_n}, \{\mathbf{x_0}, \mathbf{x_1}, \ldots, \mathbf{x_{n-1}}\}\right) \tag{3}$$

where $\mathbf{t_n}$ is the input feature maps, $\mathbf{x_n}$ is the output of the nth scale DBP module, and $\{\mathbf{x_1}, \ldots, \mathbf{x_{n-1}}\}$ are the outputs of DBP module of 1th to $(n-1)$th scales.

Specifically, F_{n} first calculates the differentiated features $\Delta\mathbf{x}_i$ between $\mathbf{t_n}$ and all its preceding scales $\{\mathbf{x_0}, \mathbf{x_1}, \ldots, \mathbf{x_{n-1}}\}$, which is defined as

$$\Delta\mathbf{x}_i = \mathrm{up}^{n-i}\left(\mathbf{t_n}\right) - \mathbf{x}_i. \tag{4}$$

As in [40], up^{n-i} denotes the projection operator which up-samples the input feature $\mathbf{t_n}$ to have the same dimension as $\mathbf{x}_i$, and $i = 0, 1, \ldots, n-1$. Then, all differentiated features and $\mathbf{t_n}$ are added to get the enhanced features of nth scale. Before performing the addition operation, $\Delta\mathbf{x}_i$ is downsampled to have the same dimension as $\mathbf{t_n}$. The formulation of these operations is defined as

$$\mathbf{x_n} = \sum_{i=0}^{n-1} \mathrm{down}^{n-i}\left(\Delta\mathbf{x}_i\right) + \mathbf{t_n} \tag{5}$$

where down^{n-i} denotes the projection operator and $n-i$ represents the factor of the downsampling operation. Finally, we get the enhanced features $\mathbf{x_n}$ as the output of the nth scale DBP module.

These enhanced features are gradually integrated into the down-sampling process, which can make up for the missing high-resolution feature information. This enables our network to extract the underlying features of the HSI more accurately. At the same time, the decoder can also directly use these enriched feature information to improve the quality of HSI reconstruction. The ablation studies shown in Section V verify the effectiveness of the proposed DBP module.

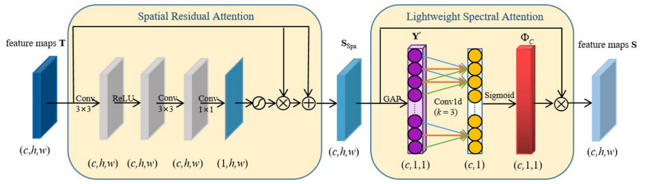


Fig. 5. Diagram of the SSA module, which consists of a spatial residual attention block and a lightweight spectral attention block. “$+$” and “$\times$” denote element-wise addition and element-wise product, respectively.

B. SSA Boosted Decoder

Corresponding to the encoder, the SSA boosted decoder also has five scales. All the scales of the decoder have the configuration of deconvolution block and residual block. For the two to zero scales, they are additionally configured with SSA modules. At the end of the decoder, a 1×1 convolutional layer with sigmoid activation function is used to make the output of the network the same channel as the original HSI and normalize the range of each item in the output to [0, 1].

The SSA module is designed to capture the coupled spatial–spectral correlation for reconstruction. As shown in Fig. 5, the SSA module concatenates the spatial residual attention block and the lightweight spectral attention block and captures the spatial and spectral correlations serially. We denote the input feature maps of the SSA module as $\mathbf{T} \in R^{c\times h\times w}$. The computation process of the SSA module can be expressed as

$$\mathbf{S} = H_{\mathrm{Spe}}\left(H_{\mathrm{Spa}}\left(\mathbf{T}\right)\right) \tag{6}$$

where $\mathbf{S} \in R^{c\times h\times w}$ is the output feature map, $H_{\mathrm{Spa}}(\cdot)$ is the function mapping of the spatial residual attention block and $H_{\mathrm{Spe}}(\cdot)$ is the function mapping of the lightweight spectral attention block.

Spatial residual attention block uses the convolution operation and residual connection to calculate spatial attention map and highlight important spatial features. This block takes the feature maps $\mathbf{T}$ as input and calculates the output $\mathbf{S}_{\mathrm{Spa}} \in R^{c\times h\times w}$ as

$$\mathbf{S}_{\mathrm{Spa}} = H_{\mathrm{Spa}}\left(\mathbf{T}\right) = \mathbf{T} + \mathbf{T} \odot \mathrm{Sigmoid}\left(\mathrm{Conv}\left(\mathbf{T}\right)\right) \tag{7}$$

where $\odot$ is the element-wise multiplication. Conv represents the convolution block. In this block, we first use there 3×3 convolutions and one 1×1 convolution to extract spatial features and obtain the spatial attention map after a sigmoid operation. The spatial attention map and input features are then multiplied and added element-wise to get the output of $\mathbf{S}_{\mathrm{Spa}}$. The spatial residual attention block can well extract the informative spatial features of HSIs.

Lightweight spectral attention block employs a local cross-channel interaction strategy to capture correlations between spectral bands. This block takes the $\mathbf{S}_{\mathrm{Spa}} \in R^{c\times h\times w}$ as input. $\mathbf{S}_{\mathrm{Spa}}$ is then processed by channel-wise global average pooling to obtain feature $\mathbf{Y}' \in R^{c\times 1\times 1}$. As in [42], we employ 1-D convolutions with kernel size k to capture local cross-channel interaction information. The kernel size k of 1-D convolution is much smaller than c, so our spectral attention block has much fewer parameters than commonly used fully connected layers.

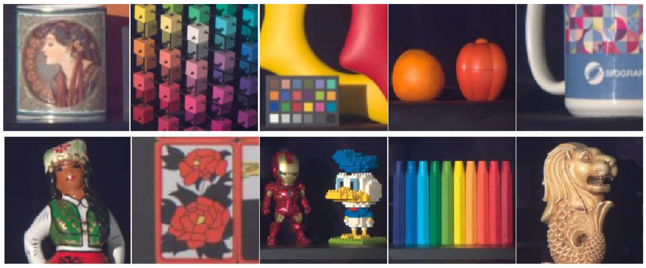


Fig. 6. Ten testing scenes used in simulation.

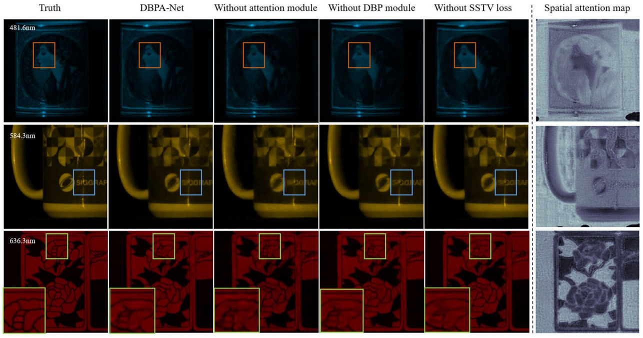


Fig. 7. Left side is a spectral band visualization of different ablation experiments of three scenes. On the right side is the space attention map. From top to bottom correspond to scenes 1, 5, and 7.

With this fast 1-D convolution, we can get the channel weight vector $\mathbf{\Phi}_c \in R^{c\times 1}$.

Specifically, the 1-D convolution of $\mathbf{Y}'$ is determined by the learnable weight matrix $\mathbf{W}_c \in R^{c\times(c+k-1)}$ and $\mathbf{W}_c$ is denoted as

$$\begin{bmatrix} w_1^1 & \cdots & w_1^k & 0 & 0 & \cdots & \cdots & 0 \\ 0 & w_2^2 & \cdots & w_2^{2+k-1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & 0 & w_i^i & \cdots & w_i^{i+k-1} & 0 & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \cdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & w_c^c & \cdots & w_c^{c+k-1} \end{bmatrix}$$

where $w_i^i, \ldots, w_i^{i+k-1}$ indicates the convolution kernel weight corresponding to k entry of $\mathbf{Y}'$. Obviously, matrix $\mathbf{W}_c$ is a sparse matrix and only $k \times c$ entries are nonzeros. The volume of weight parameters is much less than many spectral attention modules [28], [30], [43]. The weight ϕ_i corresponding to the ith feature channel is calculated as

$$\phi_i = \text{Sigmoid}\left(\sum_{j=0}^{k-1} w_i^{i+j} \mathbf{Y}'_{i+j-(k-1)/2}\right). \tag{8}$$

To avoid information loss at the boundary of $\mathbf{Y}'$, $(k-1)/2$ zeros are padded at the beginning and end of $\mathbf{Y}'$, respectively. The channel weight vector $\mathbf{\Phi}_c$ is expressed as $\mathbf{\Phi}_c = [\phi_1, \phi_2, \ldots, \phi_c]^T$.

Thus, the spectral attention module can be rewritten as

$$\mathbf{S} = H_{\text{Spe}}\left(\mathbf{S}_{\text{Spa}}\right) = \mathbf{\Phi}_c \odot \mathbf{S}_{\text{Spa}} \tag{9}$$

where $\odot$ represents the element-wise multiplication.

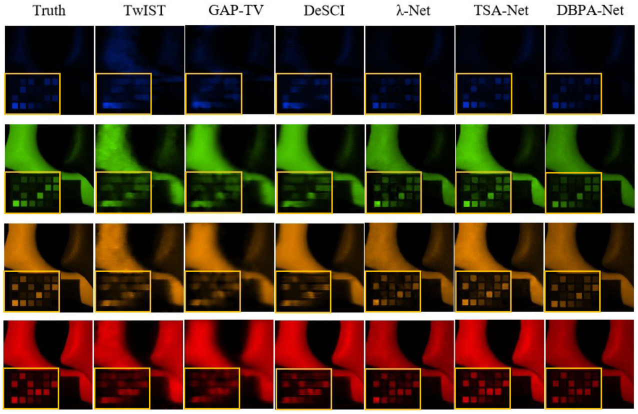


Fig. 8. Example of image reconstruction by six algorithms for scene 3. The four spectral bands are at wavelengths 462.1 nm, 551.4 nm, 594.4 nm, and 636.3 nm from top to bottom. From the left column to the rightmost column correspond to Ground truth, TwIST (PSNR21.14/SSIM0.764), GAP-TV(23.19/0.757), DeSCI(26.56/0.877), λ-Net (29.42/0.916), TSA-Net(30.03/0.921), and ours (31.34/0.938).

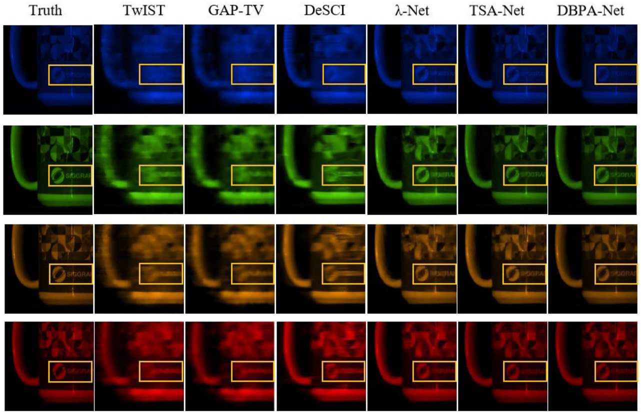


Fig. 9. Example of image reconstruction by six algorithms for scene 5. The four spectral bands are at wavelengths 462.1 nm, 551.4 nm, 594.4 nm, and 636.3 nm from top to bottom. From the left column to the rightmost column correspond to Ground truth, TwIST (PSNR21.68/SSIM0.688), GAP-TV(222.31/0.674), DeSCI(24.80/0.778), λ-Net (27.84/0.866), TSA-Net(28.89/0.878), and ours (30.02/0.924).

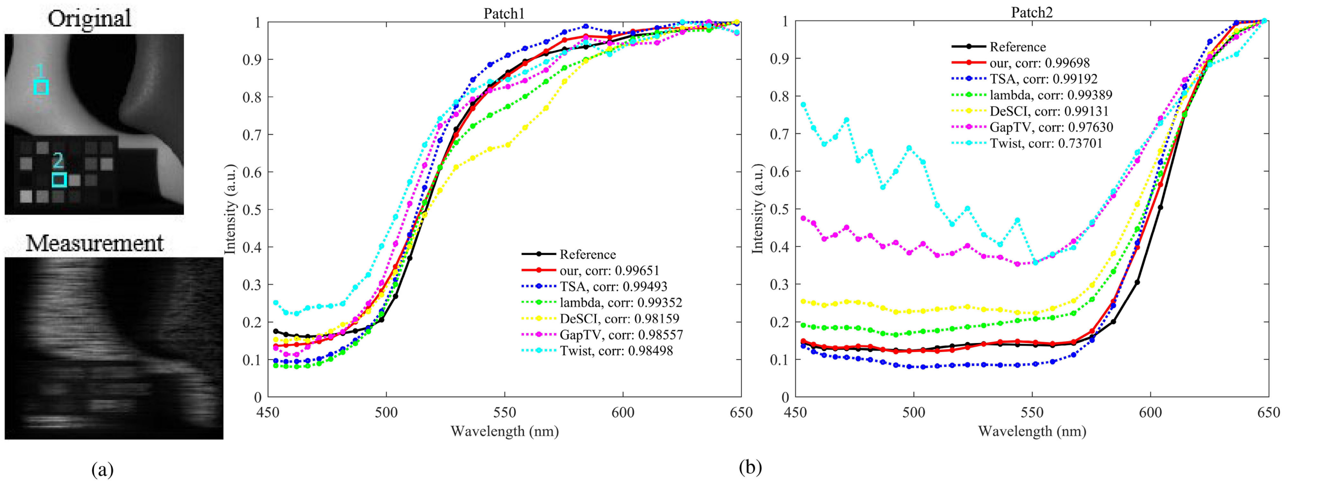
The SSA module combines a spatial residual attention block and a lightweight spectral attention block, which can effectively represent the coupled spatial–spectral correlation. Compared with other attention modules, our method achieves good reconstruction quality with minimal parameters. The ablation studies shown in Section V verify the effectiveness of the proposed SSA module.

C. Loss Function

To effectively guide our network learning, we design a compound loss function consisting of the reconstruction loss and the spatial–spectral total variation loss (SSTV), which is defined as

$$L_{\text{total}}\left(\mathbf{\Theta}\right) = L_{\text{rec}}\left(\mathbf{\Theta}, \mathbf{I}\right) + \alpha L_{\text{SSTV}}\left(\mathbf{\Theta}, \mathbf{I}\right) \tag{10}$$

where $\mathbf{\Theta}$ refers the parameter set of our network G_{Net}, and α is the parameter that tweaks the weights of these two terms. $\mathbf{I}$



(a)

(b)

Fig. 10. Snapshot measures of scene 3 and reconstructs the spectral features. (a) Spectral band of Scene 3 and its snapshot measurement. (b) Patches 1 and 2 show the reconstructed spectral curves of two regions in the scene indicated by the rectangle.

TABLE II
ABLATION STUDY OF DBP MODULE, SSA MODULE AND SSTV LOSS AS WELL AS ABLATION STUDY OF OTHER ATTENTION MODULES UNDER THE OVERALL NETWORK STRUCTURE

Ablation	Models	PSNR	SSIM	SAM
Group 1	DBPA-Net	**31.53**	**0.928**	**0.126**
	Without DBP module	30.90↓	0.911↓	0.137↑
	Without SSA module	30.51↓	0.911↓	0.152↑
	Without SSTV loss	31.04↓	0.913↓	0.128↑
Group 2	Replace SSA with TSA [30]	30.79↓	0.914↓	0.137↑
	Replace SSA with SE [34]	30.65↓	0.906↓	0.140↑
	Replace SSA with CBAM [35]	30.61↓	0.904↓	0.153↑
	Replace SSA with SA [36]	30.72↓	0.910↓	0.146↑

The significance of bold entities indicate best values.

is the training set with HSI pairs ($\mathbf{I}_{\text{gt}}, \mathbf{Y}_{\text{mea}}$), in which $\mathbf{Y}_{\text{mea}}$ is the snapshot measurement of $\mathbf{I}_{\text{gt}}$.

Specifically, L_{rec} loss calculates the mean squared error between the ground truths and the reconstructed HSIs, and it is formulated as

$$L_{\text{rec}}(\mathbf{\Theta}) = \sum_{n=1}^{N} \|\mathbf{I}_{\text{gt}}^{n} - G_{\text{Net}}(\mathbf{Y}_{\text{mea}}^{n}, \mathbf{\Theta})\|_2^2 \tag{11}$$

where $G_{\text{Net}}(\mathbf{Y}_{\text{mea}}^{n}, \mathbf{\Theta})$ is the nth reconstructed HSI by our network G_{Net}. N denotes the number of images in one training batch.

Taking into account the spatial–spectral correlation among spectral bands, we impose the SSTV loss L_{SSTV} [44] on the output $\mathbf{I}_{\text{net}}^{n}$ and define it as

$$L_{\text{SSTV}}(\mathbf{\Theta}) = \frac{1}{N}\sum_{n=1}^{N}(\|\nabla_c \mathbf{I}_{\text{net}}^{n}\|_1 + \|\nabla_h \mathbf{I}_{\text{net}}^{n}\|_1 + \|\nabla_w \mathbf{I}_{\text{net}}^{n}\|_1) \tag{12}$$

where ∇_c, ∇_h, and ∇_w are functions to compute the gradient of $\mathbf{I}_{\text{net}}^{n}$ along spectral dimension, horizontal, and vertical direction of spatial dimensions, respectively. With these two terms, the compound loss function (10) can constrain the reconstructed HSIs to approximate the ground truth with spatial–spectral consistency.

V. EXPERIMENTS

To evaluate the performance of the proposed DBPA-Net, we have conducted a series of experiments in this section, including ablation experiments and comparison experiments with several start-of-the-art algorithms, namely TwIST [19], GAP-TV [20], DeSCI [23], U-Net [45], λ-Net [28], and TSA-Net [30]. The first three are prior-driven methods, and the last three are network-driven methods. In addition to testing on the simulation dataset, we also test the proposed method on real data captured in the real world.

A. Experimental Setting

We employ Pytorch to implement the proposed network and train it from scratch by minimizing the loss function (10) with the Adam optimizer [46]. The hyper parameters of our network are set as learning rate $lr = 0.0004$, batchsize $= 4$, $k = 3$, and $\alpha = 3e - 1$. The competing methods use the code published by their authors. Our method and the competing methods all run on an NVIDIA GTX 2080Ti GPU.

The dataset we use can be downloaded from [30] and contains 28 channels with a wavelength range of 450–650nm. The wavelength of each spectral band is 453.3, 457.6, 462.1, 466.8, 471.6, 476.5, 481.6, 486.9, 492.4, 498.0, 503.9, 509.9, 516.2, 522.7, 529.5, 536.5, 543.8, 551.4, 558.6, 567.5, 575.3, 584.3, 594.4, 604.2, 614.4, 625.1, 636.3, 648.1 nm. The training dataset contains 205 HSIs with a size of $1024 \times 1024 \times 28$. During network training each epoch, 5000 data cubes with a size of $256 \times 256 \times 28$ are cut out from these training datasets for data augmentation randomly. Following the experimental strategy used in [30], this article also uses the same test set. The test data contains ten HSIs with a size of $256 \times 256 \times 28$, and the corresponding RGB image is shown in Fig. 6. These test HSIs after mask modulation move horizontally with an accumulative

TABLE III
COMPARISON OF QUANTITATIVE PSNR AND SSIM VALUES OF SEVEN METHODS UPON TEN TEST HSIs

Algorithm	TwIST	GAP-TV	DeSCI	U-Net	λ-Net	TSA-Net	ours
Scene1	24.81, 0.730	25.13, 0.724	27.15, 0.794	28.28, 0.822	30.82, 0.880	31.26, 0.887	**32.77,0.920**
Scene2	19.99, 0.632	20.67, 0.630	22.26, 0.694	24.06, 0.777	26.30, 0.846	26.88, 0.855	**29.68,0.922**
Scene3	21.14, 0.764	23.19, 0.757	26.56, 0.877	26.02, 0.857	29.42, 0.916	30.03, 0.921	**31.34,0.938**
Scene4	30.30, 0.874	35.13, 0.870	39.00, 0.965	36.33, 0.877	37.37, 0.962	39.90, 0.964	**40.04,0.972**
Scene5	21.68, 0.688	22.31, 0.674	24.80, 0.778	25.51, 0.795	27.84, 0.866	28.89, 0.878	**30.02,0.924**
Scene6	22.16, 0.660	22.90, 0.635	23.55, 0.753	27.97, 0.794	30.69, 0.886	31.30, 0.895	**32.90,0.939**
Scene7	17.71, 0.694	17.98, 0.670	20.03, 0.772	21.15, 0.799	24.20, 0.875	25.16, 0.887	**26.89,0.913**
Scene8	22.39, 0.682	23.00, 0.624	20.29, 0.740	26.83, 0.796	28.86, 0.880	29.69, 0.887	**29.86,0.909**
Scene9	21.43, 0.729	23.36, 0.717	23.98, 0.818	26.13, 0.804	29.32, 0.902	30.03, 0.903	**32.37,0.942**
Scene10	22.87, 0.595	23.70, 0.551	25.94, 0.666	25.07, 0.710	27.66, 0.843	28.32, 0.848	**29.41,0.901**
Average	22.44, 0.703	23.73, 0.683	25.86, 0.785	26.80, 0.803	29.25, 0.886	30.15, 0.893	**31.53,0.928**

The significance of bold entities indicate best values.

two-pixel step and are integrated across the spectral dimension. Then, we get measurement data with size 256×310 and combine the mask as the network input. In order to evaluate the quality of reconstructed HSIs, three evaluation metrics are used, including peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and spectral angle mapper (SAM). Larger PSNR and SSIM values and smaller SAM values indicate better reconstruction quality.

B. Ablation Studies

We primarily designed two groups of ablation studies. For the first group of ablation studies, three ablation experiments are designed to verify the effectiveness of the DBP module, SSA module, and SSTV loss, respectively. The contribution of the corresponding modules to the performance of the network is tested by removing the corresponding modules from the overall network. Table II shows the results of the ablation experiments, in which “ ↓″ and “ ↑″ highlights the indicator trends, respectively. Using the DBP module and the SSA module can improve the PSNR values by 0.63 and 1.02 dB, the SSIM values by 0.017 and 0.017, and reduce the SAM value by 0.011 and 0.026, respectively. This fully illustrates the effectiveness of the DBP module and the SSA module. Furthermore, keeping the network architecture unchanged, we test the influence of the SSTV loss term on network learning. As shown in Table II, the SSTV loss term can also boost the reconstruction performance. The left side of Fig. 7 shows the visualization results of the ablation experiment for three scenes. On the right is a visualization of the 256×256 spatial attention maps in the third SSA module of three scenes. By incorporating DBP module, SSA module, and SSTV loss, DBPA-Net can reconstruct more structures and details. The second group of ablation studies keeps the overall network structure unchanged and replaces the SSA module with TSA, SE, CBAM, and SA attention modules. As shown in Table II, our SSA module enables the overall network to obtain the best performance, which further verifies the effectiveness of our SSA module.

C. Simulation Data Results

In the simulation experiment, the coded aperture is the same as [30], which is from the real CASSI system and is used to generate snapshot measurements. Table III shows the PSNR and SSIM values of seven methods on ten test HISs. As shown in Table III, our proposed DBPA-Net is superior to other algorithms in all ten scenes. On average, the performance of DBPA-Net is 5.67 dB higher than the state-of-the-art iterative algorithm DeSCI. Meanwhile, DBPA-Net performs 4.73 dB higher in PSNR over U-Net, 2.28 dB higher over λ-Net, and 1.38 dB higher over TSA-Net. Figs. 8 and 9 show visualization results of four spectral bands of two scenes. Obviously, spatial structures reconstructed by deep neural networks are significantly clearer than iterative algorithms. Compared with λ-Net and TSA-Net methods, our method can reconstruct better details. Figs. 10 and 11 provide the snapshot measurements corresponding to the two scenes, as well as the reconstructed spectral signatures in the patches that are indicated by the rectangles. Compared with competing methods, the correlation coefficients in the legend indicate that our method can reconstruct spectral features more accurately. The spectral signature fidelity of network-driven methods is much better than prior-driven methods. Table IV gives the SAM values of the four network-driven methods. It can be seen that our method can obtain better reconstructed spectral similarity compared to competing methods.

TABLE IV
SAM VALUES OF FOUR NETWORK-DRIVEN METHODS

Algorithm	U-Net	λ-Net	TSA-Net	DBPA-Net
scene 1	0.215	0.184	0.177	**0.137**
scene 2	0.247	0.189	0.177	**0.147**
scene 3	0.182	0.130	0.128	**0.111**
scene 4	0.321	0.213	0.185	**0.122**
scene 5	0.225	0.151	0.133	**0.096**
scene 6	0.348	0.259	0.224	**0.139**
scene 7	0.198	0.143	0.131	**0.119**
scene 8	0.285	0.234	0.204	**0.145**
scene 9	0.261	0.159	0.140	**0.120**
scene 10	0.306	0.215	0.197	**0.120**
Average	0.259	0.188	0.169	**0.126**

The significance of bold entities indicate best values.

D. Real Data Results

The real data used in our experiment is captured by a hyperspectral imaging camera [30]. The original real data has a spatial size of 660×660 and 28 spectral bands. We select two real HSIs for the performance evaluation. Our proposed method

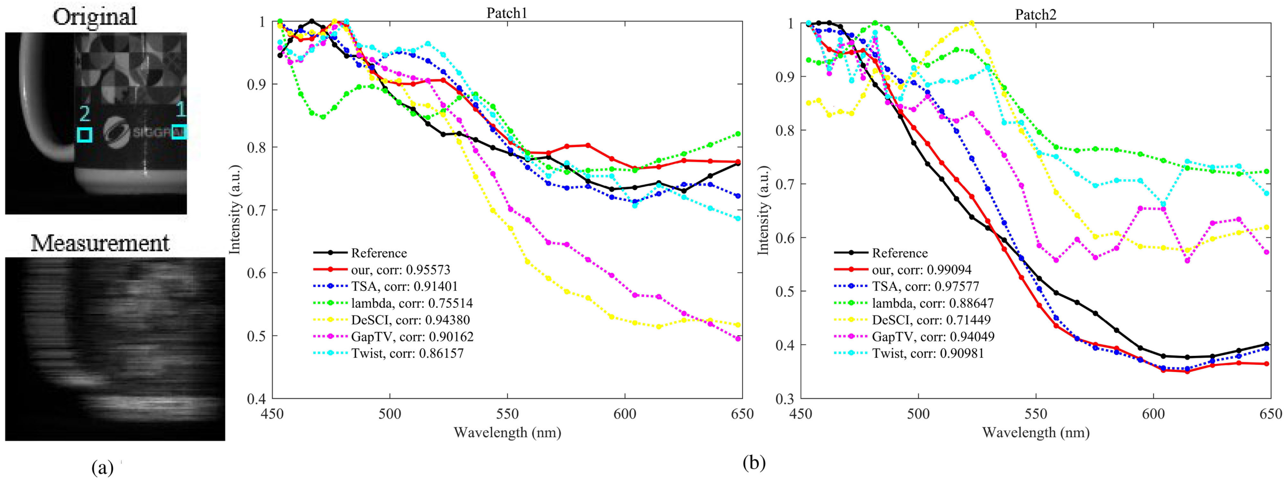


Fig. 11. Snapshot measures of scene 5 and reconstructs the spectral features. (a) Spectral band of scene 5 and its snapshot measurement. (b) Patches 1 and 2 show the reconstructed spectral curves of two regions in the scene indicated by the rectangle.

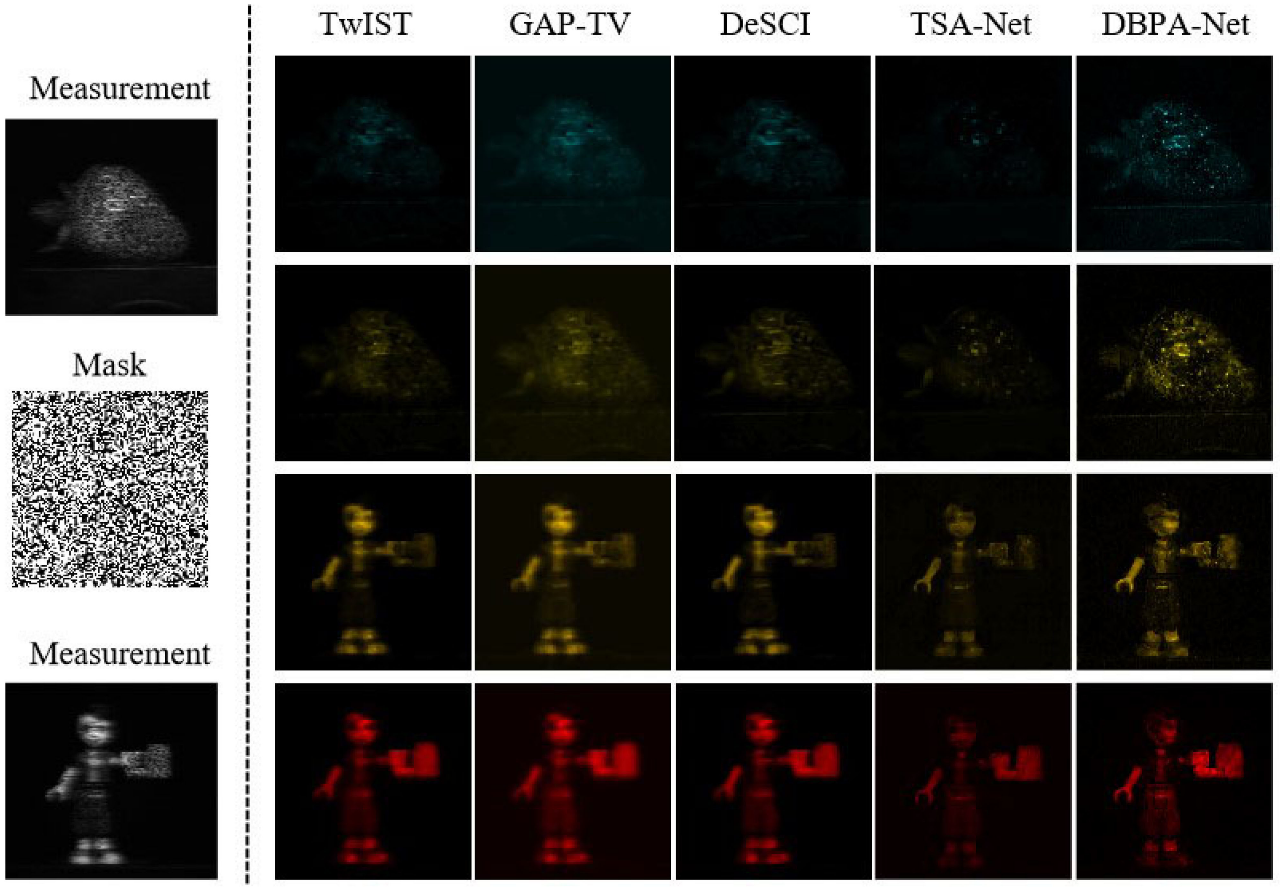


Fig. 12. Leftmost column is the snapshot measurement and coding matrix of two real data scenes $(656 \times 656 \times 28)$. On the right is the visualization of the reconstruction of each method. From the left column to the right column corresponds to TwIST, GAP-TV, DeSCI, TSA-Net, and the proposed DBPA-Net. Wavelengths from top to bottom are 486.9, 575.3, 584.3, and 636.3 nm.

was compared with TwIST, GAP-TV, DeSCI, and TSA-Net. Fig. 12 visualizes the reconstruction of four spectral bands of two scenes. As can be seen from Fig. 12, the reconstruction of TwIST and GAP-TV contains a lot of noise. Although DeSCI suppresses some noise, the reconstruction lacks texture details. The reason is that the hardware design of the real data system is complex, so the hyperspectral data snapshot measurement is easy to be disturbed by noise. This is one of the reasons why it is more difficult to reconstruct real data. Our proposed DBPA-Net can make full use of multiscale information, and retain as much high-resolution

TABLE V
AVERAGE RUN TIME FOR THE RECONSTRUCTION OF A HSI AND THE PARAMETERS VOLUME OF THREE NETWORK-DRIVEN METHODS

Algorithm	Times(s)	Parameters
DBPA-Net	**0.02**	**40.22M**
TSA-Net	0.10	44.25M
λ-Net	4.05	62.64M
DeSCI	3590.5	\
GAP-TV	19.70	\
TwIST	71.96	\

The significance of bold entities indicate best values.

spatial information as possible for reconstruction, so as to better reconstruct the detailed information. Our method achieves the best visual effect, compared with the three prior-driven methods. Compared with TSA-Net, our model provides better brightness and detail, especially strawberry leaves and Legoman's "head."

E. Time Complexity and Parameter Quantity Analysis

In addition to the quantitative evaluation of reconstruction quality, we further analyze the time complexity and the volume of parameters of the six methods. For time complexity, we compare the running time (in seconds) consumed by each method in reconstructing a single HSI with a size $256 \times 256 \times 28$. TwIST, GAP-TV, and DeSCI run on the CPU, and the other methods are trained on GPU. At the same time, we also measure the volume of parameters of the network-driven methods. Table V shows the running time results of each method and the number of parameters of partial methods. As shown in Table V, the reconstruction speed of the network-driven methods is faster than the prior-driven methods. The reason is that the network-driven methods do not require iterative optimization. The proposed DBPA-Net not only has the shortest reconstruction time, but also the least parameter volume. The DBPA-Net embodies the advantages of high reconstruction efficiency and lightweight.

VI. CONCLUSION

We propose a new lightweight DBPA-Net that can quickly reconstruct HSI from a single snapshot measurement. First, we design a DBP module to fuse complementary information between different scales for efficient reconstruction. Then, spatial and spectral correlations in HSIs are captured concisely, by employing 1-D local cross-channel connections. Furthermore, the composite loss is designed to guide network training and can reconstruct better details. Experimental results show that DBPA-Net not only exhibits better reconstruction quality than the current state-of-the-art methods, but also has the shortest reconstruction time and the least parameter volume. For the video-rate 3-D hyperspectral imaging system, it is expected that CASSI cameras combined with DBPA-Net end-to-end architecture will enjoy the benefits of rapid and high quality at the same time.

REFERENCES

[1] Y. Xu, Z. Wu, J. Li, A. Plaza, and Z. Wei, "Anomaly detection in hyperspectral images based on low-rank and sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 1990–2000, Apr. 2015.

[2] Y. Niu and B. Wang, "Extracting target spectrum for hyperspectral target detection: An adaptive weighted learning method using a self-completed background dictionary," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 1604–1617, Mar. 2017.

[3] J. Jeong et al., "Mission status of a geostationary environmental monitoring spectrometer: The development of a ground station system," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 9052–9054.

[4] C. Li, R. Hang, and B. Rasti, "EMFNet: Enhanced multisource fusion network for land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4381–4389, 2021.

[5] A. Wehr and U. Lohr, "Airborne laser scanning—An introduction and overview," *ISPRS J. Photogrammetry Remote Sens.*, vol. 54, no. 2/3, pp. 68–82, 1999.

[6] B. Fan, G. Ely, S. Aeron, and E. L. Miller, "Exploiting algebraic and structural complexity for single snapshot computed tomography hyperspectral imaging systems," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 6, pp. 990–1002, Sep. 2015.

[7] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[8] S. Bernabé, G. Martín, J. M. P. Nascimento, J. M. Bioucas-Dias, A. Plaza, and V. Silva, "Parallel hyperspectral coded aperture for compressive sensing on GPUs," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 2, pp. 932–944, Feb. 2016.

[9] Y. Xiao and J. Yang, "A fast algorithm for total variation image reconstruction from random projections," *Inverse Problems Imag.*, vol. 6, no. 3, pp. 547–563, 2010.

[10] Y. Xu, Z. Wu, J. Chanussot, M. D. Mura, A. L. Bertozzi, and Z. Wei, "Low-rank decomposition and total variation regularization of hyperspectral video sequences," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1680–1694, Mar. 2018.

[11] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[12] W.-J. Zheng, X.-L. Zhao, Y.-B. Zheng, and Z.-F. Pang, "Nonlocal patch-based fully connected tensor network decomposition for multispectral image inpainting," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8025105.

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep le arning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[14] J. Zhang and B. Ghanem, "ISTA-Net: Interpretable optimization-inspired deep network for image compressive sensing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1828–1837.

[15] 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.

[16] Y. Yang, Y. Xie, X. Chen, and Y. Sun, "Hyperspectral snapshot compressive imaging with non-local spatial-spectral residual network," *Remote. Sens.*, vol. 13, no. 9, 2021, Art. no. 1812.

[17] J. Tan, Y. Ma, H. Rueda, D. Baron, and G. R. Arce, "Compressive hyperspectral imaging via approximate message passing," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 2, pp. 389–401, Mar. 2016.

[18] M. A. T. Figueiredo, R. D. Nowak, and S. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 586–597, Dec. 2007.

[19] J. M. Bioucas-Dias and M. A. T. Figueiredo, "A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2992–3004, Dec. 2007.

[20] X. Yuan, "Generalized alternating projection based total variation minimization for compressive sensing," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 2539–2543.

[21] C. Li, W. Yin, H. Jiang, and Y. Zhang, "An efficient augmented lagrangian method with applications to total variation minimization," *Comput. Optim. Appl.*, vol. 56, no. 3, pp. 507–530, 2013.

[22] Y. Yang, J. Sun, H. Li, and Z. Xu, "Deep ADMM-Net for compressive sensing MRI," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 10–18.

[23] Y. P. Liu, X. Yuan, J. Suo, D. J. Brady, and Q. Dai, "Rank minimization for snapshot compressive imaging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 2990–3006, Dec. 2019.

[24] J. Liu, Z. Wu, L. Xiao, and X.-J. Wu, "Model inspired autoencoder for unsupervised hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5522412.

[25] D. Shen, J. Liu, Z. Wu, J. Yang, and L. Xiao, "ADMM-HFNet: A matrix decomposition-based deep approach for hyperspectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5513417.

[26] Z. Xiong, Z. Shi, H. Li, L. Wang, D. Liu, and F. Wu, "HSCNN: CNN-based hyperspectral image recovery from spectrally undersampled projections," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 518–525.

[27] I. Choi, D. S. Jeon, G. Nam, D. Gutierrez, and M. H. Kim, "High-quality hyperspectral reconstruction using a spectral prior," *ACM Trans. Graph.*, vol. 36, pp. 1–13, 2017.

[28] X. Miao, X. Yuan, Y. Pu, and V. Athitsos, "l-Net: Reconstruct hyperspectral images from a snapshot measurement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4058–4068.

[29] L. Wang, C. Sun, M. Zhang, Y. Fu, and H. Huang, "Dnu: Deep non-local unrolling for computational spectral imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1658–1668.

[30] Z. Meng, J. Ma, and X. Yuan, "End-to-end low cost compressive spectral imaging with spatial-spectral self-attention," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 187–204.

[31] T. Li, Y. Cai, Z. Cai, X. Liu, and Q. Hu, "Nonlocal band attention network for hyperspectral image band selection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3462–3474, 2021.

[32] Z. Kan, S. Li, M. Hou, L. Fang, and Y. Zhang, "Attention-based octave network for hyperspectral image denoising," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1089–1102, 2022.

[33] X. Wang and Y. Fan, "Multiscale densely connected attention network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1617–1628, 2022.

[34] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2018, pp. 7132–7141.

[35] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[36] Q.-L. Zhang and Y.-B. Yang, "Sa-net: Shuffle attention for deep convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 2235–2239.

[37] A. Wagadarikar, R. John, R. Willett, and D. Brady, "Single disperser design for coded aperture snapshot spectral imaging," *Appl. Opt.*, vol. 47, no. 10, pp. B44–B51, 2008.

[38] X. Lin, Y. Liu, J. Wu, and Q. Dai, "Spatial-spectral encoded compressive hyperspectral imaging," *ACM Trans. Graph.*, vol. 33, no. 6, pp. 1–11, 2014.

[39] M. E. Gehm, R. John, D. J. Brady, R. M. Willett, and T. J. Schulz, "Single-shot compressive spectral imaging with a dual-disperser architecture," *Opt. Exp.*, vol. 15, no. 21, pp. 14013–14027, 2007.

[40] H. Dong et al., "Multi-scale boosted dehazing network with dense feature fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2154–2164.

[41] S. Gao, M.-M. Cheng, K. Zhao, X. Zhang, M.-H. Yang, and P. H. S. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, pp. 652–662, 2021.

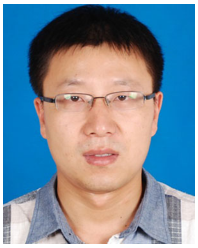
[42] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11531–11539.

[43] X. Mei et al., "Spectral-spatial attention networks for hyperspectral image classification," *Remote. Sens.*, vol. 11, 2019, Art. no. 963.

[44] H. K. Aggarwal and A. Majumdar, "Hyperspectral image denoising using spatio-spectral total variation," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, pp. 442–446, 2016.

[45] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int Conf Learn. Representations*, 2015, pp. 1–15.

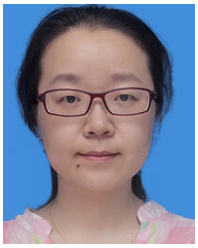


Yubao Sun received the Ph.D. degree in pattern recognition and intelligent system from the Nanjing University of Science and Technology, Nanjing, China, in 2010.

He is currently a Professor with the School of Computer and Software, Nanjing University of Information Science and Technology. His research interests are deep learning-based compressed sensing, computational imaging, video analysis, hypergraph learning, and so on.

Junru Huang received the B.S. degree in electrical engineering and automation in 2020 from the Bingjiang College, Nanjing University of Information Science and Technology, Nanjing, China, where she is currently working toward the master's degree in electronic information with the School of Automation.

Her research interests include deep learning and remote sensing image processing.



Liling Zhao received the Ph.D. degree in pattern recognition and intelligent system from the Nanjing University of Science and Technology, Nanjing, China.

Currently, she is working as an Associate Professor in the School of Automation, Nanjing University of Information Science and Technology, Nanjing, China. Her research interest covers computer vision, image recognition, and related applications in remote sensing.

Kai Hu received the Ph.D. degree in instruments science and technology from Southeast University, Nanjing, China, in 2015.

He is currently an Associate Professor and Master Tutor with the School of Automation Nanjing University of Information Science and Technology, Nanjing, China. His research interests include robot control, robot vision, and artificial intelligence.