HyperMixNet: Hyperspectral Image Reconstruction with Deep Mixed Network from a Snapshot Measurement

Kouhei Yorimoto and Xian-Hua Han
Graduate School of Science and Technology for Innovation, Yamaguchi University
Yamaguchi, 753-8511, Japan
[a037vbu, hanxhua]@yamaguchi-u.ac.jp

Abstract

Many hyperspectral imaging systems resort to computational photography technique for capturing spectral information of the dynamic world in recent decades of years. Therein, Coded aperture snapshot spectral imaging encodes the 3D hyperspectral image as a 2D compressive image (snapshot) and then employs an inverse optimization algorithm embedded in the imaging system to reconstruct the underlying HSI. This study proposes a novel HyperMixNet to reconstruct an underlying HSI from the single snapshot image. Specifically, to reduce the size of the reconstruction model for being handy embedded in the real imaging system, we integrate the MixConv block instead of the conventional convolutional layers in our proposed HyperMixNet, which can not only greatly decrease the network parameter amount but also learn multi-level context for more representative feature extraction. Simultaneously, we employ a mixed spatial and spectral convolutional module to effectively learn the spatial structure and spectral attribute for more robust HSI reconstruction. We further design a mixed loss function for network training, which incorporates not only spatial fidelity but also spectral fidelity aiming at recovering the hyperspectral signature with small spectral distortion. Experimental results on three benchmark HSI datasets validate that our proposed method outperforms the state-of-the-art methods in quantitative values, visual effect, and reconstruction model scale.

1. Introduction

Hyperspectral imaging captures the detail spectral distribution of a scene as a three-dimensional cubic image, and can describe spectral intensities at decades or hundreds of wavelength bands of each pixel location. The rich spectral information in the hyperspectral image (HSI) greatly benefits the characterization of the imaged scene and has been widely used to various fields ranging from remote sensing

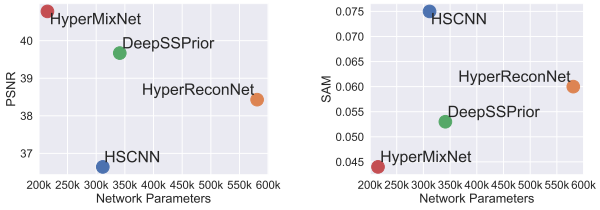


Figure 1: HSI reconstruction performance and model size comparison of deep learning based methods. Our method exploits multiple MixSS modules for hierarchically reconstructing spatial and spectral residual components, the MixConv inside MixSS for exploring multi-level spatial contexts and parameter reduction, and the mixed loss for restoring reliable spectral information, achieving the best spatial and spectral restoring performances according to the PSNR and Sam while requiring much less parameters. Larger PSNR means better spatial fidelity index while smaller Sam denotes better spectral fidelity measurement.

[13, 5], medical diagnosis [4, 18], vision inspection to digital forensics, to name a few. For obtaining the 3D cubic data, hyperspectral imaging systems have to measure exposures at multiple times for different wavelength bands using 1D or 2D sensors, and thus it would take a long time for the imaging procedure of the scene. This limits its utilization for imaging dynamics objects and scenes or capturing video with high-speed rates [3, 21]. Therefore, several computational spectral imaging prototypes have been evolved to capture the spectral signatures of the dynamics world [9, 17, 11, 6], and inspired by the fundamental compressive sensing theory, coded aperture snapshot spectral imaging (CASSI) [24, 2, 12] has made significant progress for being prospected to capture high spatial and temporal resolution HSIs. CASSI systems are generally divided into two phases: the exposure measure phase for encoding the 3D HSI into a single 2D compressive (a snapshot) image, and the computational reconstruction phase for recovering

the underlying HSI from the snapshot measurement via employing an inverse optimization strategy. So far, various efforts have been made on both phases, and the bottleneck of the existed CASSI system mainly lies in the limited reconstruction quality of the underlying HSI. In this study, we concentrate on exploring the computational reconstruction method from a snapshot measurement in CASSI systems.

Since the reconstruction of the 3D HSI from a single snapshot image is an ill-posed problem in nature, existing methods generally leverage hand-crafted priors to regularize the inverse model for robust reconstruction. Many priors have been investigated, including total variation (TV) [15, 27], sparsity representation [28, 24], non-local similarity [10, 29] for modeling different specific characterization of the underlying HSIs, and show some improvements regard to the reconstruction performance. The widely explored priors are designed empirically and often insufficient to handle the wide variety of spectra of natural hyperspectral images, which unavoidably results in poor spectral reconstruction. With the great success of the deep learning in computer vision applications, deep convolutional neural network (DCNN) [32, 8, 19, 30, 25] has been employed for HSI reconstruction via automatically learning the underlying priors of the latent HSI, and been proven to provide much better reconstruction performance and faster reconstruction in test phase than the conventional optimization-based methods. However, researchers are making efforts to design more complicated and deeper network for boosting reconstruction performance, which would result in large-scale reconstruction model and be difficult to be implanted into the real hyperspectral imaging systems. Moreover, the current DCNN based methods usually explore the reconstruction errors such as the mean squared error of the prediction and the ground-truth HSI as the loss function for network training, which mainly measure the spatial fidelity of the reconstruction. However, in HSI reconstruction scenario, the spectral characteristic preservation is more essential than the spatial detail maintenance. Further, there are still some rooms for performance improvement in the HSI reconstruction field.

To handle the above mentioned limitations, this study proposes a novel deep mixed neural network for hyperspectral image reconstruction (HyperMixNet). The proposed HyperMixNet employs mixed spatial and spectral convolutional modules (MixSS modules) to effectively learn the spatial structure and spectral attribute for more robust HSI reconstruction. In each MixSS module, we leverage the MixConv block inside a single layer, which consists of different groups of Depthwise convolutional layers with various kernel sizes, for spatial structure reconstruction and follow a spectral block for spectral correlation exploration and spectral attribute recovering. Specifically, the exploited MixConv block can not only greatly reduce the network parameters but also simultaneously learn the representative features with multiple contexts for adaptively context mixture. Moreover, to reconstruct more reliable spectral signature, we combine spectral and spatial fidelity measure together to formulate the mixed loss function for the HyperMixNet training. Our proposed HyperMixNet is verified on three representative hyperspectral image datasets, i.e., CAVE [34], Harvard [7] and ICVL [1], and outperforms the state-of-the-art methods in quantitative values, visual effect, and reconstruction model scale. Figure 1 manifests the performance and model size comparisons with the state-of-the-art deep learning based methods. In summary, our main contributions are three-fold:

1. We present a novel HyperMixNet for HSI reconstruction from a single snapshot measurement, which employs multiple mixed spatial and spectral convolutional modules for reliable spatial structure and spectral characteristic reconstruction.

2. We leverage the MixConv block inside a single layer for both model parameter reduction and multiple context fusion.

3. We exploit spectral and spatial fidelity measure to construct mixed loss function for network training.

2. Related Work

Recently, the hyperspectral reconstruction models in the computational spectral imaging have been actively researched, which are mainly divided into two directions: optimization-based methods and deep learning-based methods, and substantial improvements have been witnessed. In this section, we briefly survey the related work.

2.1. Optimization-based methods

Due to the ill-posed nature of the HSI reconstruction inverse problem from a snapshot measurement, previous methods have striven to explore different hand-crafted priors for modeling the spatial structure and spectral characteristics of the latent HSI, and then regularize the inverse model for robust optimization. Via considering the high dimensionality of spectral signatures, many previous approaches usually exploit the image local priors for characterizing the spectral image structure within a local region. Wang et al. [27] proposed to impose the first-order smoothness prior via regularizing the image gradients and resulted in the Total Variation (TV) regularized model to spectral image reconstruction. Further Yuan et. al. [35] employed generalized alternating projection (GAP-TV) while Kittle et. al. [15] explored two-step iterative shrinkage/thresholding method (TwIST). Utilizing TV prior for the HSI reconstruction generally benefits both boundary preservation and

smooth region recovery but may lead to the detail structure lost. Recently motivated by blind compressed sensing (BCS) [20], sparse representation methods have been applied for solving the CASSI reconstruction problem, which learns dictionary to model the sparsity prior to image patch [17]. Later, Yuan et. al. employed the compressibility constraint rather than sparse representation prior and proposed to learn the dictionary via leveraging global-local shrinkage prior [36] while Wang et. al. [29] investigated 3D non-local sparse representation model via integrating non-local similarity for boosting reconstruction performance. However, the hand-crafted image priors are not always sufficient to capture the characteristics in various spectral images, and to discover a proper prior for a specific scene is still hard task in the real scenario.

2.2. Deep learning-based methods

Deep learning-based methods can effectively learn the complex and high-representative features containing different contexts and have been widely applied in HSI processing. In HSI reconstruction scenario, instead of carefully designing priors for modeling the spatial and spectral characteristics of the latent HSI, deep learning-based methods aim at implicitly learning the prior from the previously prepared training samples, and then constructing the mapping model between the compressive image and the desirable HSI. Xiong et al [32] initially exploited a CNN-based hyperspectral image recovering method (HSCNN) from spectrally under-sampled projections and evaluated the feasibility for HSI reconstruction from a common RGB image or a compressive sensing (CS) measurement. Wang et al. [30] explored a joint coded aperture optimization and image reconstruction from compressive HS imaging for automatically learning the optimal sensing matrix and reconstructing the latent HSI under the learned coded aperture in an end-to-end manner. Moreover, Miao et al. [19] developed a dual-stage generative model, dubbed as λ-net, for hierarchically reconstructing the HIS while Wang et al. [25] proposed an end-to-end CNN network to learn multi-stage deep spatial-spectral priors (DSSP) for modeling both local coherence and dynamic characteristics of the underlying HSI. Although promising performance has been achieved with the deep network, the recent research line mainly focuses on more complicated and deeper network architecture for performance boosting, which generally leads to large-scale reconstruction model. However, the large scale of the offline learned model would limit wide applicability for being implanted in real hyperspectral imaging systems.

Recently, to increase the flexibility of deep reconstruction model, several works integrated deep learned priors into iterative optimization procedure and developed some deep unrolling based optimization methods in natural compressive sensing, e.g., LISTA [14] ADMMNet [19, 33] and ISTA-Net [37]. Choi et al. [8] proposed to learn spectral prior via employing the convolutional auto-encoder and integrated the learned deep image priors in pretraining step into the optimization procedure as a regularizer. Wang et al. [26] further exploited both spectral and non-local (NLS) prior learning, and integrated the NLS-based regularization into the model-based optimization method for robust HSI reconstruction in spectral compressive sensing. However, the deep unrolled method still requires to conduct the mathematical optimization under the regularization by the deep learned priors, which would be time-consuming in the HSI reconstruction phase. Moreover, all existing methods for HSI reconstruction including optimization-based and deep learning-based methods explore the reconstruction errors such as the mean squared error of the prediction and the ground-truth HSI as criteria for optimization or network training, which mainly measure the spatial fidelity of the reconstruction. In HSI reconstruction scenario, the spectral characteristic preservation is more essential than the spatial detail maintenance, and thereby spectral fidelity criteria should be exploited to facilitate the reliable spectral reconstruction of the latent HSI.

3. The Proposed HyperMixNet for HSI Reconstruction

CASSI [24, 2, 12] encodes the 3D hyperspectral information into a 2D compressive image. The incident light for a spectral scene: $X(h, w, \lambda)$, where h and w are the spatial index ($1 \leqq h \leqq H$, $1 \leqq w \leqq W$) and λ is the spectral index ($1 \leqq \lambda \leqq \Lambda$), is collected by the objective lens, and spatially modulated with a transmission function $T(h, w)$ created in coded aperture coded. Then the modulated scene is spectrally dispersed with a wavelength-dependent dispersion function $\psi(\lambda)$ by the disperser, and follows the charge-coupled device (CCD) for detecting the spatial and spectral coded scene as a snapshot image. The observation model for the 2D snapshot image can be formulated as:

$$Y(h, w) = \sum T(h - \psi(\lambda))X(h - \psi(\lambda), w, \lambda). \quad (1)$$

The goal of the HSI reconstruction in the CASSI is to reconstruct the underlying 3D spectral image $\mathbf{X}$ from the compressed measurement $\mathbf{Y}$. This study proposes a novel HyperMixNet for effective and efficiently HSI Reconstruction.

3.1. Overview

The schematic concept of the proposed HyperMixNet is shown in Figure 2, which includes the reconstruction moduole and multiple mixed spatial and spectral convolutional modules (MixSS) for hierarchically reconstructing the unrecovered residual spatial and spectral components. The MixSS module consists of reciprocal spatial convolutional

block for exploiting the spatial correlation in local regions and spectral convolutional blocks for exploring correlation among all spectral channels, which is also dubbed as intermixing (InterMix) among convolutional layers. Moreover, due to the large variety of objects in the imaged scene, the spectra of different spatial positions should have various sizes of correlated regions, and thus the spatial correlation exploration in a fixed size of regions would not achieve adaptive reconstruction for all spatial positions. We advocate to employ a MixConv spatial block, which includes several groups of Depthwise convolutional layers with various kernel sizes inside one layer, and also dubbed as intramixing (IntraMix). The integration of the IntraMix block not only significantly decreases the network parameters for easily being implanted in the HS imaging system but also simultaneously explores the spatial correlation of different sizes of regions inside one layer. Then the following spectral block concurrently exploit spectral correlation and conduct deep mixture of different feature maps for the multiple spatial contexts. Furthermore, we combine spectral and spatial fidelity measure together to formulate loss function, and propose a mix loss function (MixLoss) for the HyperMixNet training. Next, we describe in detail the backbone architecture of the proposed HyperMixNet, different modules in it and the proposed MixLoss.

3.2. HyperMixNet

The proposed HyperMixNet mainly comprises two types of modules: the initial reconstruction module and multiple MixSS modules for hierarchically refining spatial and spectral reconstruction. Given the observed snapshot image Y, our goal is to reconstruct the full spectral image X using the HyperMixNet. Firstly, the initial reconstruction module, which includes several plain convolutional layers, converts the 2D compressed image Y into the initial HSI: $\mathbf{X}^{(0)}$ via expanding the dimension in the spectral direction from 1 to Λ. The initial reconstruction can be formulated as:

$$\mathbf{X}^{(0)} = f_{recon}(\mathbf{Y}), \tag{2}$$

where $f_{recon}(\cdot)$ denotes the transformation in the initial reconstruction module. In our experiments, we simply adopt 3 convolutional layers with kernel size 3×3 following a RELU activation after each layer. After that, we stack multiple mixed spatial and spectral convolutional modules (MixSS) and employ residual connection to construct the backbone architecture for hierarchically reconstructing more fine-grained spatial structure and spectral characteristics in the latent HSI. Let $\mathbf{X}^{(k)}$ denotes the output of the $k-th$ MixSS module, the output of the $(k+1)-th$ MixSS module can be expressed as

$$\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} + f_{MixSS}(\mathbf{X}^{(k)}), \tag{3}$$

where $f_{MixSS}(\cdot)$ denotes the transformation operators in the MixSS module. The MixSS module consists of spatial and spectral reconstruction blocks for exploring the correlation in both directions, and is expected to reconstruct more reliable structures in both directions. Moreover, we employ the residual connection in the MixSS module to learn only the un-recovered components in the previous module as shown in Figure 2. Next, we would describe the detail architecture in the MixSS module.

3.3. The MixSS Module

In the HSI reconstruction scenario from a snapshot image, it is needed to simultaneously recover the detail structure in spatial directions and the specific fine signature in the spectral directions. It is a challenge task to employ a plain network architecture for reconstructing the high-dimensional signal in both directions. In this study, we propose a mixed spatial and spectral reconstruction module (MixSS) to reciprocally exploit the correlation in a local spatial region and the spectral correlation in all channels, and stack multiple MixSS modules for hierarchically reconstructing the un-recovered residual spatial and spectral components. The architecture of the MixSS module is shown in Figure 2 which includes the spatial block for local spatial correlation exploration, an optional spatial context selection (SCS) block and the spectral block for band correlation probing. Next, we describes the detail operations in the three blocks.

Spatial conv block: Given the intermediately reconstructed HSI $\mathbf{X}^{(k)} \in \mathbb{R}^{H\times W\times \Lambda}$ at the $k-th$ module, the MixSS module first employs the spatial convolutional block to learn high representative features via taking account of the spatial context, which consists of a vanilla convolutional layer following a RELU layer and a mixed depth-wise convolutional layer (dubbed as MixConv) [23] for parameter reduction and adaptive spatial correlation exploration in different sizes of local spatial regions. The spatial block is formulated as:

$$\mathbf{X}^{(k)}_{spatial} = f_{MixConv}(f_{Con-Relu}(\mathbf{X}^{(k)})), \tag{4}$$

where $f_{Con-Relu}(\cdot)$ denotes the transform operations of the first vanilla convolutional layer with the spatial kernel size 3×3 to expand the channel dimension from Λ to L and the followed RELU layer while $f_{MixConv}(\cdot)$ represents the operations of the mixed depth-wise convolutional layer. Let's denote the intermediate outputted feature map of the spatial block as $\bar{\mathbf{X}}^{(k)} = f_{Con-Relu}(\mathbf{X}^{(k)})$, and then concretely formulate the mathematical transformation of the MixConv block. We partition the intermediate feature map $\bar{\mathbf{X}}^{(k)} \in \mathbb{R}^{H\times W\times L}$ into M groups: $\bar{\mathbf{X}}^{(k)} = [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_M]$ via evenly dividing the channel dimension, where $\mathbf{X}_m \in \mathbb{R}^{H\times W\times L_m}$ ($L_m = L/M$) represents the feature maps in the $m-th$ group. The MixConv layer is employed to ex-

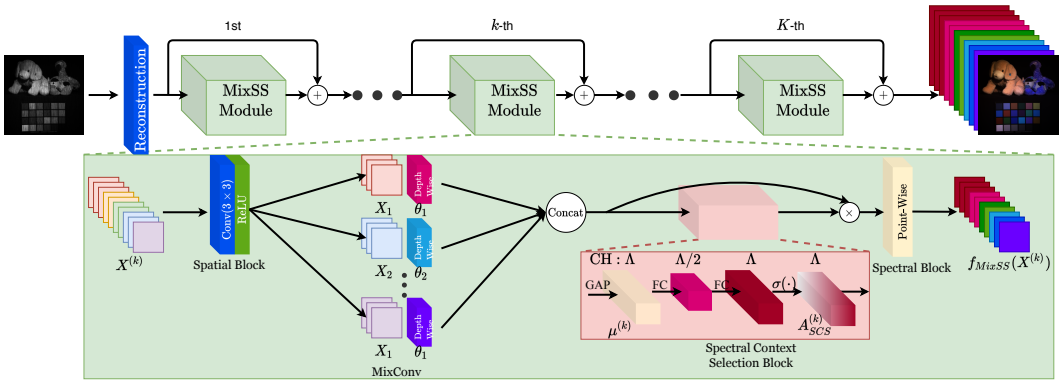


Figure 2: The architecture of the proposed HyperMixNet consisting of multiple MixSS modules, and the detail illustration of the MixSS module.

ploit different spatial contexts for different groups via using depth-wise convolutional layers. Let's denote the parameter set of the MixConv layer as $\mathbf{\Theta}_{Mix}^{(k)} = \{\theta_1, \theta_2, \cdots, \theta_M\}$ in the M group of depth-wise convalutional layers, where the parameters for different groups have various spatial kernel sizes for exploring spatial contexts in different local regions with $\theta_m \in \mathbb{R}^{s_m \times s_m \times L_m}$, the MixConv layer is formulated as:

$$\mathbf{X}_{spatial}^{(k)} = Concat(f_{dp}^{\theta_1}(\mathbf{X}_1), f_{dp}^{\theta_2}(\mathbf{X}_2), \cdots, f_{dp}^{\theta_M}(\mathbf{X}_M)), \tag{5}$$

where $f_{dp}^{\theta_m}(\cdot)$ represents the depth-wise convolutional layer with the weight parameter θ_m (for simplicity, we ignore the bias parameters). With the different kernel spatial sizes at different groups, the spatial correlation in various local regions is simultaneously integrated for extracting high representative features in one layer. Moreover, we employ the depth-wise convolutional operations in all groups, which can greatly reduces the parameters ($\frac{1}{L_m}$) compared with a vanilla convolutional laye for being easily implanted in the real imaging systems, and expect more reliable spatial structure reconstruction via concentrating on spatial context exploration.

Spatial context selection block: The output of the spatial conv block includes M groups of feature maps possessing different spatial contexts, which would contributes variously to the HSI reconstruction according to the content of the latent HSI. This study investigates a plug and play block, dubbed as spatial context selection (SCS) block, for adaptively learning the contribution index of the spatial feature maps, and optionally plug it inside the MixSS module to form another version: HyperMixNet$_+$. In the SCS block, we firstly aggregate the feature maps with different spatial contexts via employing global average pooling to generate the global contribution indexes for all channels:

$$\mu_{ch}^{k} = \frac{1}{W \times W} \sum_{w=1}^{H} \sum_{h=1}^{H} x_{spatial,ch}^{(k)}(w, h) \tag{6}$$

where $x_{spatial,ch}^{(k)}$ denotes the excitation value on the spatial position (w, h) and the $ch-th$ channel of the spatial feature map $\mathbf{X}_{spatial}^{(k)}$. With the global contribution index vector $\mu^k = [\mu_1^k, \mu_2^k, \cdots, \mu_L^k]$ abtained using Eqn. (6), we further explore the channel correlation for capturing the channel-wise dependencies via two fully connected (FC) layers, and then follows a non-linear transformation using an activation function to generate the adaptive spatial context selection (SCS) vector, expressed as:

$$A_{SCS}^{(k)} = \sigma(f_{FC}(f_{FC}(\mu^k))) \tag{7}$$

where $f_{FC}(\cdot)$ denote the transformation operation of a fully-connected layer, and $\sigma(\cdot)$ represents the sigmoid activation function. Finally, the SCS vector are combined with the raw spatial feature map for automatically emphasizing the representative feature maps with important spatial contexts and attenuating the channels with irrelevant spatial contexts for HSI reconstruction, formulated as:

$$\bar{\mathbf{X}}_{spatial}^{(k)} = A_{SCS}^{(k)} \otimes \mathbf{X}_{spatial}^{(k)} \tag{8}$$

where $\otimes$ is element-wise multiplication, and the elements of $A_{SCS}^{(k)}$ are automatically replicated in the horizontal and vertical directions of the spatial domain to match the dimensions of $\mathbf{X}_{spatial}^{(k)}$. It should note that the SCS block reduces the spatial feature map to $L-$dimensional global vector for learning the contribution factor for different spatial contexts, and needs very little additional parameters for network learning.

Spectral conv block: With the spatial feature maps by the spatial conv block or the SCS block, we aim at reconstructing the $(k+1)-th$ stage of HSI via aggregating $\mathbf{X}_{spatial}^{(k)}$ or $\bar{\mathbf{X}}_{spatial}^{(k)}$ into a Λ-band of cubic data. A

point-wise convolutional layer is employed to fuse the features with different spatial contexts, and is also expected to exploit the correlation among different channels (spectral bands). The fusion and exploration of the channel correlation is formulated as:

$$\bar{\mathbf{X}}^{(k+1)} = f_{PW}(\bar{\mathbf{X}}^{(k)}_{spatial}) \tag{9}$$

where $f_{PW}(\cdot)$ denotes the transformation operation of the point-wise convolutional layer. Since the point-wise convolutional layer focuses on only the correlation exploration of different spectral channels without considering the spatial correlation, we expect more reliable spectral reconstruction. Moreover the point-wise convolutional layer has much less parameters than a vanilla conv layer for benefiting our small-scale reconstruction model.

3.4. The mixed loss function

In HSI reconstruction scenario, spectral recovering fidelity would greatly affect the performance of the downstream tasks in the hyperspectral analysis systems. Thus the spectral characteristic preservation is more essential than the spatial detail maintenance. The existing deep models for HSI reconstruction usually explore the reconstruction errors such as the mean squared error of the prediction and the ground-truth HSI as the loss function for network training. The loss function with the reconstruction error mainly measure the spatial fidelity while completely overlooking the spectral fidelity, and thus generally results in high spectral distortion. To handle the limited spectral recovering problem, this study combines spectral fidelity, where the spectral angle mapper (SAM) [16] metric is used for evaluating the spectral reliability, and spatial fidelity (the conventional mean square error) to formulate a mixed loss function (MixLoss) for network training. Given the $n-th$ groundthruth HSI and its corresponding prediction from our HyperMixNet: $\mathbf{X}_n$ and $\hat{\mathbf{X}}_n$, the SAM value between $\mathbf{X}_n$ and $\hat{\mathbf{X}}_n$ is computed as:

$$L_{SAM} = \frac{1}{W \times H} \sum_{w=1}^{W} \sum_{h=1}^{H} \frac{(\mathbf{x}(w,h), \hat{\mathbf{x}}(w,h))}{\|(\mathbf{x}(w,h))\|_2 \cdot \|(\hat{\mathbf{x}}(w,h))\|_2} \tag{10}$$

where w and h represent the pixel positions in vertical and horizontal directions, and $(\cdot,\cdot)$ denotes the dot product of two vectors. Let denotes the MSE loss as L_{MSE}, the MixLoss is formulated as,

$$L_{MixLoss}(\mathbf{X}_n, \hat{\mathbf{X}}_n) = \alpha r_{MSE} L_{MSE} + (1-\alpha) r_{SAM} L_{SAM} \tag{11}$$

where r_{MSE} and r_{SAM} are scale-adjusting parameters for normalizing two losses to the same order ($r_{MSE} = 1$ and $r_{SAM} = 0.1$ in our experiments) while α ($0 \leqq \alpha \leqq 1$) is a hyper-parameter for balancing the contribution between the two losses.

4. Experiment and Result

4.1. Experiment setting

To demonstrate the effectiveness of our proposed HypeMixNet model, we conduct extensive experiments on three benchmark hyperspectral datasets including the CAVE dataset [34], the Harvard dataset [7], and the ICVL dataset [1]. The CAVE dataset consists of 32 images with spatial resolution 512 $\times$ 512 while the Harvard dataset is composed of 50 outdoor images with spatial resolution 1040 $\times$ 1392 captured under daylight conditions. The ICVL dataset, which is by far the most comprehensive HSI dataset with large variety of natural scenes, is composed of 201 images with spatial resolution 1300 $\times$ 1392, from which we select a subset with 104 images as the dataset in our experiments. All HSI images in the three datasets have 31 spectral channels ranging from 420nm to 720nm for the CAVE and from 400nm to 700nm for the Harvard and ICVL datasets. We randomly select 22 images in CAVE dataset, 40 images in Harvard dataset, and 90 images in ICVL dataset for network training and the remainder for testing. For simulating the 2D snapshot image, we construct the transmission function $T(h,w)$ of the coded aperture in the HS imaging system via randomly generating a binary matrix according to a Bernoulli distribution with p = 0.5, and then transform the original HSI with the transformation function to the snapshot measurements. To prepare samples for network training, we extract the corresponding patches with spatial size of 48 $\times$ 48 from the original HSIs and their snapshot measurements. We compare our proposed method with several state-of-the-art HSI reconstruction methods, including four traditional methods with handcrafted prior modeling , i.e, TwIST with TV prior [15, 27], AMP with sparsity prior [22], and 3DNSR and SSLR with NLS prior [29, 10], and three deep learning-based methods, i.e., HSCNN [32], HyperReconNet [30], and Deep Spatial Spectral Prior (DeepSSPrior) [25]. To evaluate the effect of integrated MixSS module in HyperMixNet, we experimentally vary the number K of the MixSS module with 5, 7, and 9, respectively, and provide the final HSI reconstruction for comparison. We also set the value of α in MixLoss to 0.5 for combining spatial and spectral fidelity losses, and please refer to the supplemental material for more compared results for various values of α. Three quantitative metrics including peak signal-to-noise (PSNR), structural similarity (SSIM) [31], and spectral angle mapping (SAM) [16] are employed to evaluate the performance of different HSI reconstruction methods. PSNR and SSIM measure the spatial fidelity of the reconstructed HSIs, which are calculated on each 2D spatial image, and averaged over all spectral bands. SAM assesses the spectral fidelity, which is calculated on each 1D spectral vector and averages over all spatial points. Larger values of PSNR and SSIM suggest better

Table 1: Performance comparisons on ICVL, Harvard and CAVE datasets (3% compressive ratio). The best performance is labeled in bold, and the second best is labeled in underline.

Dataset	Metrics	TwIST	AMP	3DNSR	SSLR	HSCNN	HyperReconNet	DeepSSPrior	Our($K=5$)	Our($K=7$)	Our($K=9$)
ICVL	PSNR	26.15	24.56	27.95	29.16	36.64	38.43	39.67	39.19	<u>40.16</u>	**40.70**
	SSIM	0.936	0.909	0.958	0.964	0.963	0.972	0.979	0.976	<u>0.980</u>	**0.981**
	SAM	0.053	0.09	0.051	<u>0.046</u>	0.075	0.060	0.053	0.052	0.047	**0.044**
Harvard	PSNR	27.16	24.96	28.51	29.68	35.09	36.04	37.62	36.81	**37.72**	<u>37.70</u>
	SSIM	0.924	0.935	0.94	0.952	0.936	0.938	**0.955**	0.947	<u>0.952</u>	0.951
	SAM	0.119	0.155	0.132	<u>0.101</u>	**0.092**	0.166	0.130	0.127	0.119	0.120
CAVE	PSNR	(−)	(−)	(−)	(−)	23.22	<u>25.82</u>	24.82	25.57	**25.88**	25.81
	SSIM	(−)	(−)	(−)	(−)	0.720	<u>0.829</u>	0.807	0.818	0.814	**0.831**
	SAM	(−)	(−)	(−)	(−)	0.475	0.305	0.392	0.290	**0.259**	<u>0.260</u>
Paramaeters		(−)	(−)	(−)	(−)	311,541	580,709	341,173	120,017	167,503	214,989

Table 2: Ablation results of the proposed HyperMixNet and HyperMixNet$_+$ on all three datasets.

Dataset	Compare Model	HyperMixNet								HyperMixNet$_+$							
	Group number M	1		2		3		4		1		2		3		4	
	Loss	MSE	Mix	MSE	Mix	MSE	Mix	MSE	Mix	MSE	Mix	MSE	Mix	MSE	Mix	MSE	Mix
ICVL	PSNR	39.36	39.24	39.96	39.84	40.34	40.24	39.36	40.70	37.13	38.01	38.13	37.81	38.83	38.72	38.86	39.12
	SSIM	0.977	0.977	0.979	0.979	0.981	0.980	0.977	0.981	0.967	0.976	0.977	0.976	0.979	0.977	0.980	0.979
	SAM	0.055	0.051	0.049	0.049	0.048	0.046	0.055	0.044	0.061	0.063	0.054	0.053	0.050	0.051	0.056	0.050
Harvard	PSNR	37.56	37.61	37.1	37.22	37.22	37.47	37.28	37.7	36.88	36.73	37.37	36.9	37.33	37.16	36.98	36.87
	SSIM	0.953	0.953	0.952	0.951	0.946	0.950	0.948	0.951	0.949	0.945	0.953	0.948	0.955	0.951	0.949	0.949
	SAM	0.131	0.126	0.144	0.124	0.143	0.124	0.134	0.12	0.157	0.133	0.135	0.131	0.136	0.125	0.153	0.131
CAVE	PSNR	25.62	24.26	25.56	25.75	25.69	25.94	25.85	25.81	24.24	23.90	24.50	24.56	24.01	24.95	24.42	25.14
	SSIM	0.818	0.783	0.802	0.813	0.825	0.838	0.837	0.831	0.808	0.809	0.817	0.789	0.823	0.813	0.823	0.831
	SAM	0.342	0.320	0.348	0.268	0.333	0.263	0.324	0.260	0.394	0.312	0.361	0.293	0.360	0.280	0.364	0.268

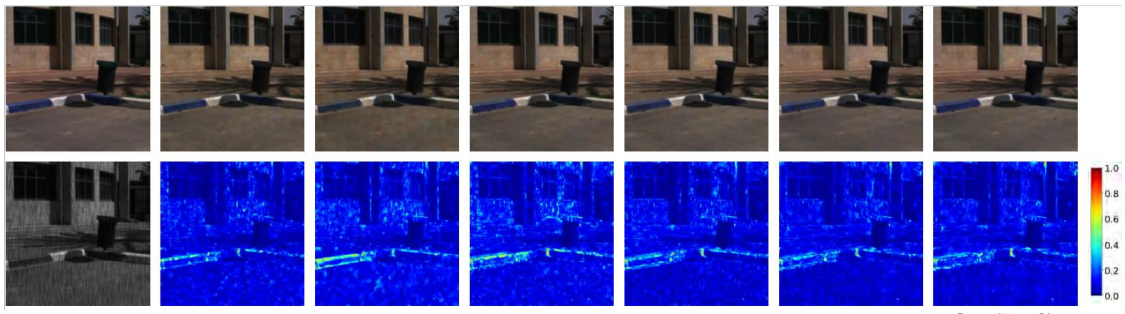
performance, while a smaller value of SAM implies a better reconstruction.

4.2. Numerical Results

Table 1 manifests the compared quantitative evaluation of the reconstructed HSIs on ICVL, Harvard, and CAVE datasets using both traditional methods and deep learning-based methods including our proposed HyperMixNet, and the network parameters to be learned in deep learning-based methods. Moreover, since our proposed HyperMixNet is mainly composed of a serial of MixSS modules, where the module number can be adapted according to the trade-off of the reconstruction effectiveness and efficiency, we also provide the compared results via varying the module numbers in Table 1. From Table 1, it can be seen that the proposed HyperMixNet outperforms most state-of-the-art methods in both spatial and spectral fidelity indexes excepting a little drop of SAM value on the Harvard dataset compared with the SSLR and HSCNN methods. However, the number of parameters in our proposed HyperMixNet with $K = 9$ MixSS modules is reduced by 31%, 63%, and 37% compared with HSCNN, HyperReconNet, and DeepSSPrior, respectively. In addition, the network parameter of our HyperMixNet can be further reduced with $K = 7$ but without largely affecting the reconstruction performances for all three datasets. Thus, it can be concluded that our proposed HyperMixNet not only reduces the reconstruction model size but also simultaneously improves the HSI restoration accuracy.

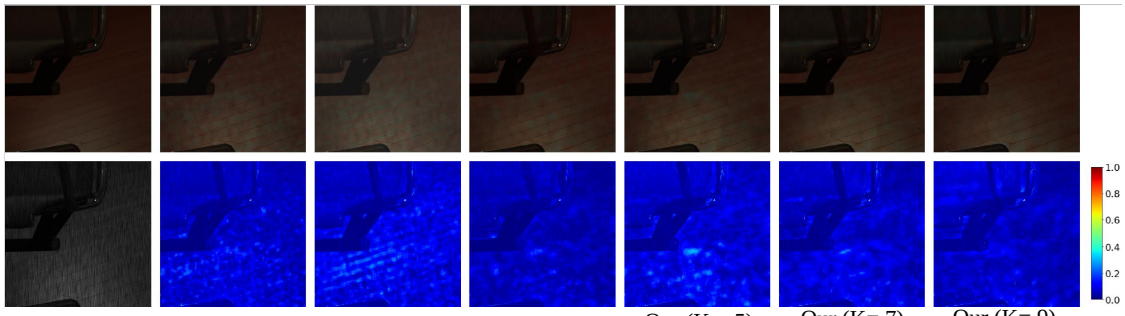
4.3. Ablation Studies

In this section, we present the ablation study for exploring different integrated components in the MixSS module, and the used losses: MSE loss and the proposed mixed loss for the network training. As described in subsection 3.3, we exploited the MixConv layer for reconstructing spatial structure, which consists of M groups to explore different level spatial contexts (the standard depth-wise convolutional layer with $M = 1$), and integrate the SCSB block to investigate another version, dubbed as HyperMixNet$_+$, for adaptively selecting useful spatial features. We conducted experiments with different group numbers ($1 \leq M \leq 4$) of the MixConv layer, MSE/Mixed losses (α = 0.5 for Mixed loss. Please referrer to more ablation results with different α values) on both HyperMixNet and HyperMixNet$_+$. Table 2 gives the ablation results on all three datasets. From Table 2, we can see that larger group number in the MixConv layer manifests better results on all datasets, and the mixed loss provides large improvement in spectral fidelity (SAM value) while simultaneously maintaining or improving spatial fidelity (PSNR and SSIM values). The explored HyperMixNet$_+$ with the optional SCSB leads to no performance boosting compared to the HyperMixNet while it achieves comparable performance with the benchmark deep learning-based methods. In the future, we are going to explore more effective selecting module for spatial contexts.



GT / Compressive | HSCNN (37.51, 0.047) | HyperReconNet (35.66, 0.063) | DeepSSPrior (39.29, 0.040) | Our (K = 5) (38.84, 0.040) | Our (K= 7) (39.97, 0.037) | Our (K= 9) (40.40, 0.034)

Figure 3: Visual quality comparison of the representative reconstructed image of HSCNN [32], HyperReconNet [30], DeepSSPrior [25] and our proposed HyperMixNet (K =5, 7, 9) from ICVL dataset. The PSNR and SAM for the resulted images are shown in the parenthesis.



GT / Compressive | HSCNN (37.46, 0.095) | HyperReconNet (35.41, 0.126) | DeepSSPrior (39.81, 0.076) | Our (K = 5) (38.03, 0.087) | Our (K= 7) (38.96, 0.080) | Our (K= 9) (40.08, 0.075)

Figure 4: Visual quality comparison of the representative reconstructed image of HSCNN [32], HyperReconNet [30], DeepSSPrior [25] and our proposed HyperMixNet (K =5, 7, 9) from Harvard dataset. The PSNR and SAM for the resulted images are shown in the parenthesis.

4.4. Perceptural Quality

To visualize reconstruction results, two representative images from the ICVL and the Harvard datasets using HSCNN, HyperReconNet, DeepSSPrior, and our proposed HyperMixNet ($K = 5, 7, 9$) are shown in Figure 3 and Figure 4. The first column shows the compressive snapshot image by CASSI and its corresponding Ground-Truth image while the other columns in the first and second rows provide the reconstructed images and the absolute difference images between the ground-truth and the reconstruction using different deep learning-based methods. To conduct visualization, which considers the intensities of all spectral bands, we convert the HSIs to sRGB space via the CIE color matching function. From Figure 3, we can see that the proposed HyperMixNet is able to reconstruct wall colors and the building, which cannot be reconstructed by conventional deep learning-based models, much closer to the ground truth. Figure 4 also shows that the reconstructed HSI with the proposed HyperMixNet ($K = 9$) achieved the best spatial and spectral restoration results.

5. Conclusion

This study proposed a more efficient and effective deep model, named as HyperMixNet, for the HSI reconstruction. HyperMixNet mainly consists of multiple MixSS modules for reciprocally exploring spatial and spectral correlations and hierarchically restoring the spatial and spectral residual components. Specifically, to reduce parameters and exploit different levels of spatial context, we adopted a MixConv layer for implementing the spatial reconstruction layer. Moreover, to recover more reliable spectral information of the latent HSI, we integrated spectral fidelity measure into the loss function of network training, and proposed a mixed loss for accounting for both spatial and spectral restoration degrees. Experimental results on three HSI datasets showed comparable or better performance than the state-of-the-art methods, and much less parameters than the benchmark deep models.

References

[1] Boaz Arad and Ohad Ben-Shahar. Sparse recovery of hyperspectral signal from natural rgb images. In *ECCV*, 2016.

[2] G. Arce, D. Brady, L. Carin, H. Arguello, and D. Kittle. Compressive coded aperture spectral imaging: An introduction. *IEEE Signal Processing Magazine*, 31:105–115, 2014.

[3] R. Basedow, D. C. Carmer, and M. Anderson. Hydice system: implementation and performance. In *Defense, Security, and Sensing*, 1995.

[4] Asgeir Bjorgan and L. L. Randeberg. Towards real-time medical diagnostics using hyperspectral imaging technology. In *European Conference on Biomedical Optics*, 2015.

[5] Marcus Borengasser, William S. Hungate, and R. Watkins. Hyperspectral remote sensing: Principles and applications. 2007.

[6] Xun Cao, Hao Du, Xin Tong, Q. Dai, and Stephen Lin. A prism-mask system for multispectral video acquisition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:2423–2435, 2011.

[7] Ayan Chakrabarti and Todd E. Zickler. Statistics of real-world hyperspectral images. *CVPR 2011*, pages 193–200, 2011.

[8] Inchang Choi, Daniel S. Jeon, Giljoo Nam, Diego Gutierrez, and Min H. Kim. High-quality hyperspectral reconstruction using a spectral prior. *ACM Transactions on Graphics (TOG)*, 36:1 – 13, 2017.

[9] Qi Cui, Jongchan Park, R. Theodore Smith, and L. Gao. Snapshot hyperspectral light field imaging using image mapping spectrometry. *Optics letters*, 45 3:772–775, 2020.

[10] Y. Fu, Yinqiang Zheng, I. Sato, and Y. Sato. Exploiting spectral-spatial correlation for coded hyperspectral image restoration. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3727–3736, 2016.

[11] Liang Gao, R. Kester, N. Hagen, and T. Tkaczyk. Snapshot image mapping spectrometer (ims) with high sampling density for hyperspectral microscopy. *Optics Express*, 18:14330 – 14344, 2010.

[12] M. Gehm, R. John, D. Brady, R. Willett, and T. Schulz. Single-shot compressive spectral imaging with a dual-disperser architecture. *Optics express*, 15 21:14013–27, 2007.

[13] A. Goetz, G. Vane, J. Solomon, and B. Rock. Imaging spectrometry for earth remote sensing. *Science*, 228:1147 – 1153, 1985.

[14] K. Gregor and Y. LeCun. Learning fast approximations of sparse coding. In *ICML*, 2010.

[15] David S. Kittle, Kerkil Choi, Ashwin A. Wagadarikar, and David J. Brady. Multiframe image estimation for coded aperture snapshot spectral imagers. *Applied optics*, 49 36:6824–33, 2010.

[16] Fred A. Kruse, A. B. Lefkoff, Joseph W. Boardman, Kathleen B. Heidebrecht, A. T. Shapiro, P. J. Barloon, and Alexander F. H. Goetz. The spectral image processing system (sips) interactive visualization and analysis of imaging spectrometer data. 1993.

[17] Xing Lin, Yebin Liu, J. Wu, and Q. Dai. Spatial-spectral encoded compressive hyperspectral imaging. *ACM Trans. Graph.*, 33:233:1–233:11, 2014.

[18] Guolan Lu and B. Fei. Medical hyperspectral imaging: a review. *Journal of Biomedical Optics*, 19, 2014.

[19] Xin Miao, Xin Yuan, Yunchen Pu, and Vassilis Athitsos. lambda-net: Reconstruct hyperspectral images from a snapshot measurement. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4058–4068, 2019.

[20] Ajit Rajwade, David S. Kittle, Tsung-Han Tsai, David J. Brady, and Lawrence Carin. Coded hyperspectral imaging and blind compressive sensing. *SIAM J. Imaging Sciences*, 6:782–812, 2013.

[21] Y. Schechner and S. Nayar. Generalized mosaicing: Wide field of view multispectral imaging. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24:1334–1348, 2002.

[22] Jin Tan, Y. Ma, H. Rueda, D. Baron, and G. Arce. Compressive hyperspectral imaging via approximate message passing. *IEEE Journal of Selected Topics in Signal Processing*, 10:389–401, 2016.

[23] M. Tan and Quoc V. Le. Mixconv: Mixed depthwise convolutional kernels. In *BMVC*, 2019.

[24] Ashwin A. Wagadarikar, Renu John, Rebecca M. Willett, and David J. Brady. Single disperser design for coded aperture snapshot spectral imaging. *Applied optics*, 47 10:B44–51, 2008.

[25] Laibao Wang, Chen Sun, Ying Fu, Min Hoi Kim, and Hua Huang. Hyperspectral image reconstruction using a deep spatial-spectral prior. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8024–8033, 2019.

[26] Lizhi Wang, Chen Sun, M. Zhang, Y. Fu, and Hua Huang. Dnu: Deep non-local unrolling for computational spectral imaging. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1658–1668, 2020.

[27] Lizhi Wang, Zhiwei Xiong, Dahua Gao, Guangming Shi, and Feng Juan Wu. Dual-camera design for coded aperture snapshot spectral imaging. *Applied optics*, 54 4:848–58, 2015.

[28] Lizhi Wang, Zhiwei Xiong, Dahua Gao, Guangming Shi, Wenjun Zeng, and Feng Wu. High-speed hyperspectral video acquisition with a dual-camera architecture. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4942–4950, 2015.

[29] L. Wang, Z. Xiong, G. Shi, Feng Wu, and Wenjun Zeng. Adaptive nonlocal sparse representation for dual-camera compressive hyperspectral imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:2104–2111, 2017.

[30] Lizhi Wang, Tao Zhang, Ying Fu, and Hua Huang. Hyperreconnet: Joint coded aperture optimization and image reconstruction for compressive hyperspectral imaging. *IEEE Transactions on Image Processing*, 28:2257–2270, 2019.

[31] Zhengjiang. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004.

[32] Zhiwei Xiong, Zhan Shi, Huiqun Li, Lizhi Wang, Dong Liu, and Feng Wu. Hscnn: Cnn-based hyperspectral image recovery from spectrally undersampled projections. *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 518–525, 2017.

[33] Y. Yang, J. Sun, Huibin Li, and Zongben Xu. Admm-csnet: A deep learning approach for image compressive sensing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:521–538, 2020.

[34] Fumihito Yasuma, T. Mitsunaga, Daisuke Iso, and S. Nayar. Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum. *IEEE Transactions on Image Processing*, 19:2241–2253, 2010.

[35] Xin Yuan. Generalized alternating projection based total variation minimization for compressive sensing. *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2539–2543, 2016.

[36] X. Yuan, T. Tsai, Ruoyu Zhu, P. Llull, D. Brady, and L. Carin. Compressive hyperspectral imaging with side information. *IEEE Journal of Selected Topics in Signal Processing*, 9:964–976, 2015.

[37] J. Zhang and Bernard Ghanem. Ista-net: Interpretable optimization-inspired deep network for image compressive sensing. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1828–1837, 2018.