Learning Texture Enhancement Prior with Deep Unfolding Network for Snapshot Compressive Imaging*

Mengying Jin[0000−0001−5582−1015], Zhihui Wei[0000−0002−4841−6051], and Liang Xiao[0000−0003−0178−9384]

Nanjing University of Science and Technology, Nanjing 210094, China
{jinmengying_maths,gswei}@njust.edu.cn
xiaoliang@mail.njust.edu.cn

Abstract. Coded Aperture Snapshot Spectral Imaging (CASSI) utilizes a two-dimensional (2D) detector to capture three-dimensional (3D) data, significantly reducing the acquisition cost of hyperspectral images. However, such an ill-posed problem desires a reliable decoding algorithm with a well-designed prior term. This paper proposes a decoding model with a learnable prior term for snapshot compressive imaging. We expand the inference obtained by Half Quadratic Splitting (HQS) to construct our Texture Enhancement Prior learning network, TEP-net. Considering the high-frequency information representing the texture can effectively enhance the reconstruction quality. We then propose the residual Shuffled Multi-spectral Channel Attention(Shuffled-MCA) module to learn information corresponding to different frequency components by introducing the Discrete Cosine Transform (DCT) bases. In order to overcome the drawbacks of grouping operations within the MCA module efficiently, we employ the channel shuffle operation instead of a channel-wise operation. Channel shuffle rearranges the channel descriptors, allowing for better extraction of channel correlations subsequently. The experimental results show that our method outperforms the existing state-of-the-art method in numerical indicators. At the same time, the visualization results also show our superior performance in texture enhancement.

Keywords: Coded Aperture Snapshot Spectral Imaging (CASSI) · deep unfolding · residual shuffled multi-spectral channel attention · texture enhancement · channel shuffle.

1 Introduction

Hyperspectral data has been used in a wide range of applications, including agriculture[12], vegetation and water resource studies [6], surveillance [33] and

* This research was funded by the National Natural Science Foundation of China under Grant 61871226; in part by the Fundamental Research Funds for the Central Universities under Grant NO. JSGP202204; in part by the Jiangsu Provincial Social Developing Project under Grant BE2018727.

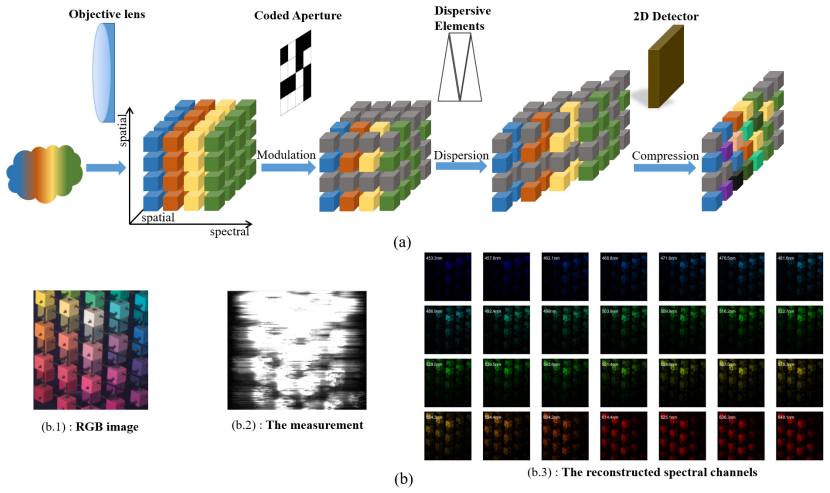


Fig. 1. (a). The data flow of the Coded Aperture Snapshot Spectral Imaging (CASSI) system. (b.1) The RGB image ; (b.2) The measurement; (b.3) The reconstructed spectral channels. Color for better view.

so on. However, the high spectral resolution of hyperspectral data makes it a certain difficulty in acquisition and storage. Inspired by the theory of compressed sensing, compressed imaging system was developed and is now widely employed for high-speed videos and hyperspectral images. Snapshot Compressive Imaging (SCI) systems combine compressed sensing with optical sensors, i.e., compressing multiple frames of data into a single snapshot measurement, offering the advantages of low cost, low bandwidth, and high speed, and becoming a popular compressed imaging system. As one of the representative SCI systems for hyperspectral image acquisition, the Coded Aperture Snapshot Spectral Imaging (CASSI) [5, 22, 23] system is of tremendous research value.

The compressed imaging systems always rely on a corresponding decoding algorithm to obtain the original data, and CASSI is no exception. As depicted in Fig.1, the 3D data is captured by the objective lens first, then modulated by the coded aperture, dispersed by the dispersion element, and finally overlaps on the 2D detector plane. Thereby, it is plagued by several issues that cause the inverse problem under the SCI task to be quite challenging. The conventional model-based computational imaging methods mimic the physical process of imaging and incorporate the prior term in order to compress the solution space. In the selection of the prior items, TwIST[2] utilizes the total variation, whereas DeSCI[11] employs patch-based weighted nuclear norm. These approaches[2, 11, 32, 29] use hand-crafted priors derived from domain knowledge, mainly focusing on extracting generic features, lacking the application of features for the target data itself.

With the development of Convolutional Neural Network(CNN), the CNN-based methods[28, 18, 16, 3, 17] use the powerful learning ability of neural networks to learn features from the data, substantially improving the results. However, they are also being accused of being uninterpretable. Recently, Zhang et al.[34] experimentally demonstrate that neural networks have the capability to express arbitrarily complex functions, which allows CNN to be considered as a learnable proximal operator. This makes it intuitive to think of viewing a deep learning network as a feature prior learning module[37] and combining it with conventional model-based approaches to construct a novel unfolding network for addressing the SCI problem.

Although researchers[34] have proved that CNN can learn arbitrary patterns, including high-frequency information[24], they also confirm that the CNNs tend to learn '*Simple*' patterns, i.e., low-frequency information, first. Redundant low-frequency information may affect the propagation of high-frequency information[1]. So treating channels of the input feature equally obviously cannot be a reliable solution. The channel attention[9] offers a workaround by applying channel-wise re-weighting to the input features, which means we can use the channel attention mechanism to enhance the target information that we need selectively.

As stated before, the two-dimensional data acquired by the CASSI system is radially ambiguous, with a significant quantity of missing data. The previous technologies mainly focused on how to recover incomplete data and did not care about the texture information lost, resulting in a lack of clarity. The DGSMP[10] is concerned about this and proposes using a deep unfolding Gaussian Scale Mixed model with spatial adaptation to learn edge texture information and obtain state-of-the-art results.

Motivated by these, this paper will construct an unfolding network for processing SCI tasks with novel texture-enhancement prior learning term. **Firstly**, considering the significance of the texture represented by the high-frequency information for the image recovery effect and the ability of the channel attention mechanism to enhance the information selectively, we introduce a Multi-spectral Channel Attention(MCA) Mechanism for texture learning. **Secondly**, the conventional MCA would assign the same Discrete Cosine Transform(DCT) base to multiple frames in each group after grouping channels. Instead of assigning DCT bases frame by frame[15] resulting in a significant increase in computation, we prefer to introduce a simple and improved method, i.e., channel shuffle. Then, we incorporate the MCA with channel shuffle into a basic residual structure to build the complete texture enhancement prior learning module, so-called the residual Shuffled MCA or SMCA for simplicity. The skipping connections in the residual structure ensure that low-frequency information can propagate backward, allowing the main network to focus on enhancing high-frequency information. **Finally**, we unfolded the data fidelity term into a network form, combined it with the constructed SMCA, and treated it as one stage. Repeating the stage several times will form a complete network. We did a series of experiments and

discovered that we received state-of-the-art results on several metrics, with the visualization demonstrating that the texture was successfully enhanced.

The contribution of our work can be summarised as follows:

1. We propose a novel unfolding network with texture enhancement prior learning module for SCI tasks.
2. We introduce MCA to selectively enhance the information corresponding to the different frequencies.
3. We improve the drawback in MCA caused by grouping operation by introducing channel shuffle and coupling the residual structure to construct the Shuffled MCA.
4. Experiments demonstrate the superior performance of our method compared to state-of-the-art methods in terms of numerical evaluation metrics and visualization results.

2 Related Works

Previous Work. As there are arbitrarily many 3D images $\mathcal{X}$ that can be reduced to the same 2D image y, Snapshot Compressive Imaging (SCI) is an ill-conditioned problem. In order to choose the most plausible result among the candidate set, a proper regularization is required. Conventional model-based methods usually use hand-crafted priors, such as TwIST[2] and GAP-TV[32] uses total variation regularization, GMM-online[29] uses Gaussian Mixture Model, DeSCI[11] uses weighted nuclear norm, etc. Although they all have complete theoretical proof but also rely on artificial parameter tuning, while time-consuming, DeSCI requires even hours to process one single image. Neural networks have shown great potential in recent years, with much of the CNN-based works taking recovery to a new level[28, 18, 16, 3, 17]. However, the uninterpretability of CNNs is undesirable given the existence of a complete physical imaging mechanism in computational imaging.

Gregor et al.[7] propose to unfold the inference of conventional model-based methods into neural networks, transforming the methods into parameters learnable while still being interpretable and significantly reducing the time cost. Istanet[36] and ADMM-net[30] expand the computational flow graph directly into a network and update the parameters efficiently by back propagation. Ma et al.[13] propose Tensor ADMM-net, which further relates the iterative steps with the neural network operations one to one. Zhang et al.[37] then consider using CNN as the prior term of the network. In this paper, we also try to employ CNN, which can fit arbitrary functions, construct the prior learning module, and consider it a learnable proximal operator.

Channel Attention Mechanism. CNNs have achieved impressive outcomes, but they usually treat the channels of the input features equally and ignore the correlation between channels. Squeeze-and-Excitation Network(SENet)[9] believes it is better to directly and explicitly model the dynamic nonlinear dependencies between channels. SENet uses the Global Average Pooling(GAP)

operation to 'Squeeze' the information. Otherwise, using the mean value as the feature descriptor has proved that only the lowest frequency information of the image is focused on, while high-frequency information is out of consideration. Thus, CBAM[27] and BAM[19] combine global maximum pooling and GAP to extract richer information. Recently, FcaNet[20] has considered placing the "Squeeze" work under the frequency domain by grouping the input features and then convolving them with the pre-computable 2D DCT bases. They also prove mathematically that when and only when the lowest frequency component in DCT bases is selected, it will be equivalent to the GAP operation.

Channel Shuffle. Grouping operation[35, 26] has been widely used in deep learning, effectively processing different components adaptively with less computation. However, grouping operation still leads to information imbalance between groups. In order to address this, ShuffleNet[39, 14] introduced the channel shuffle operation to rearrange the channel order after the group convolution, which effectively promotes information fusion between groups. CPN[21] proposed introducing channel shuffle in multi-scale cascaded pyramid networks to enhance the multi-scale information fusion across channels. SA-Net[38] applies the spatial-spectral joint attention mechanism to the grouped data separately and improves the balance between efficiency and performance by applying the channel shuffle operation over the concatenate feature.

3 Proposed Method

3.1 Reconstruction Model with Learnable Prior for CASSI

We assume the expected 3D spectral information is $\mathcal{X} \in \mathbb{R}^{M\times N\times\lambda}$, the 2D measurement captured by the CASSI system is the $y \in \mathbb{R}^{M\times(N+k(\lambda-1))}$, where M and N represent spatial size, the λ is the spectral number, and the k represents the dispersion coefficient caused by the dispersion elements, then we have:

$$y = \Phi\mathcal{X} + n \tag{1}$$

where, $\Phi \in \mathbb{R}^{M\times(N+k(\lambda-1))\times\lambda}$ is the feed-forward response function describing the physical process in CASSI. n represents the noise.

The reconstruction model corresponding to the Eq (1) can be written as:

$$\min_{\mathcal{X}} ||y - \Phi\mathcal{X}||_F^2 + \gamma\beta(\mathcal{X}) \tag{2}$$

where $\beta(\cdot)$ is the regularization term and γ is the balance parameter. The equation (2) can be written as an unconstrained optimization problem according to the Half Quadratic Splitting (HQS) method:

$$\min_{\mathcal{X},\mathcal{Z}} ||y - \Phi\mathcal{X}||_F^2 + \eta||\mathcal{X} - \mathcal{Z}||_F^2 + \gamma\beta(\mathcal{Z}) \tag{3}$$

where η is the penalty parameter, $\mathcal{Z}$ is the auxiliary variable. The solution to the equation (3) can be split into the following two sub-problems.

$$\mathcal{X}^{k+1} = arg\min_{\mathcal{X}} ||y - \Phi\mathcal{X}||_F^2 + \eta||\mathcal{X} - \mathcal{Z}^k||_F^2 \tag{4}$$

$$\mathcal{Z}^{k+1} = arg\min_{\mathcal{Z}} \eta||\mathcal{X}^{k+1} - \mathcal{Z}||_F^2 + \gamma\beta(\mathcal{Z}) \tag{5}$$

$\mathcal{X}$ **sub-problem**: fixed variable $\mathcal{Z}$, updated variable $\mathcal{X}$. Equation (4) is a least squares problem, and the solution can be given directly in closed form as:

$$\mathcal{X}^{k+1} = (\Phi^T\Phi + \eta\mathcal{I})^{-1}(\Phi^T y + \eta\mathcal{Z}^k) \tag{6}$$

However, it is computationally expensive to calculate the inverse of $\Phi^T\Phi$ directly, so gradient descent is employed here to find the approximate solution of (6) :

$$\mathcal{X}^{k+1} = (1 - \epsilon\eta)\mathcal{X}^k - \epsilon\Phi^T\Phi\mathcal{X}^k + \epsilon\Phi^T y + \epsilon\eta\mathcal{Z}^k \tag{7}$$

where ϵ represents the step size in the gradient descent.

$\mathcal{Z}$ **sub-problem**: fixed variable $\mathcal{X}$, updated variable $\mathcal{Z}$. The formula (5) is a prior term, which can be expressed as a learnable proximal operator:

$$\mathcal{Z}^{k+1} = \mathcal{P}(\mathcal{X}^{k+1}) \tag{8}$$

To solve the Eq(3) more effectively, we are attempting to convert the model-based method into a learning-based method by extending the inference into a deep neural network. This will allow the method to learn targeted features from the data and apply them to improve performance. In other words, each iteration is viewed as a sub-module of the network, and K consecutive modules are connected in sequence to build the unfolding network. The following section will introduce the proposed texture enhancement prior learning network.

3.2 Texture Enhancement Prior Learning Network: TEP-net

TEP-net consists of three modules, the data fidelity module, the texture enhancement prior learning module (TEP-module), the system forward response and its inverse process learning module. In order to better extract the information, we also proposed a novel residual Shuffled Multi-spectral Channel Attention embedded in the TEP-module, called Shuffled-MCA. Fig.2 demonstrates the construction of the TEP-net.

Learning System Forward Response and Its Inverse. Φ and Φ^T describe the physical imaging processing and its inverse of the CASSI system, which is particularly important in CASSI. Since the network involves multiple multiplication calculations, considering the high computational complexity of tensor multiplication and the fact that the coded aperture varies in different systems and settings. In order to reduce computational complexity and increase the robustness of the system, we follow the design of [10] and use the cascaded residual

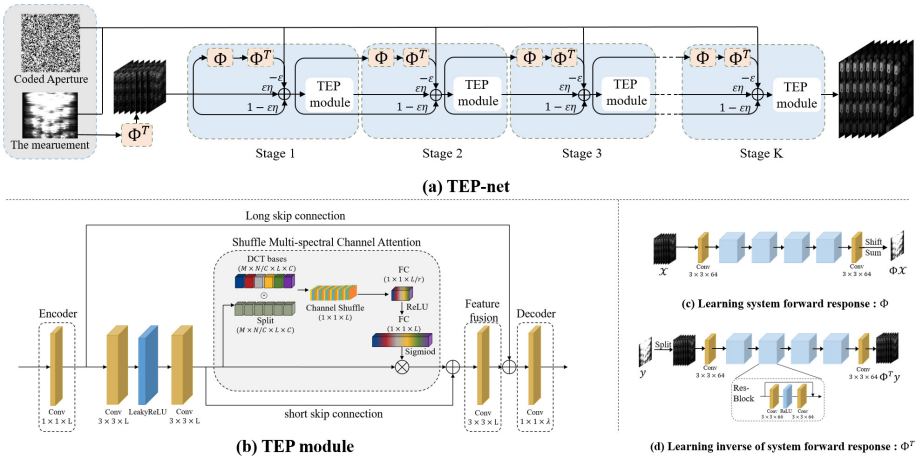


Fig. 2. The structure of the proposed method. (a): The structure of TEP-net. (b): Our proposed residual Shuffled Multi-spectral Channel Attention module for texture enhancement prior learning. (c-d): Structure of the sub-network for learning system forward response and its inverse process: Φ & Φ^T. (c) for learning Φ and (d) for learning Φ^T.

sub-network to learn Φ, Φ^T. It is easy to notice from Fig.2(a) that $\Phi^T y$ repeatedly occurs in the model. Since the coded aperture Φ_{orig} is known in the actual experiments, to balance performance and effectiveness, we use $\Phi_{orig}^T y$ as input to the model without engaging in additional learning (the top line in Fig.2(a)).

The Data Fidelity Module. As illustrated in Fig.2(a), we present this part via a data flow diagram. The network inputs are the measurement y and the $\Phi_{orig}^T y$ reconstructed using the original coded aperture Φ_{orig}. $\Phi_{orig}^T y$ will repeatedly feed into the network as the top line of Fig.2(a) shows. The measurement y will pass through a randomly initialized Φ^T learning sub-network first. Then, the obtained values $\Phi^T y$ become the initial value of the network.

Learning Texture Enhancement Prior. Conventional hand-crafted priors benefit from domain knowledge with complete theoretical analysis but lack information mining from the target data. Meanwhile, CNNs can fit arbitrary functions and learn from data, allowing CNNs to be viewed as learnable proximal operators. Thus we propose a residual shuffled multi-spectral channel attention sub-network for texture enhancement prior learning, which concatenates a feature extraction term and a shuffled multi-spectral channel attention module with two skip connections.

As shown in Fig.2(b), the input feature $\mathcal{X}^k \in \mathbb{R}^{M\times N\times \lambda}$ is first encoded by the encoder which is a point-wise convolution with L channels.

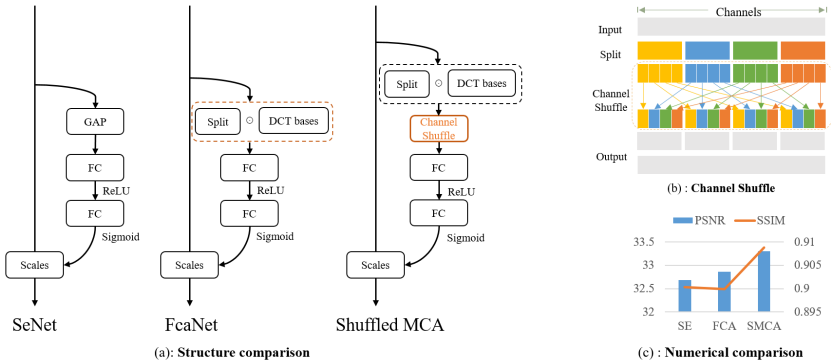


Fig. 3. The comparison of the SENet, FcaNet and the proposed Shuffle-MCA.

$$encoder(\cdot) = conv(\cdot) \tag{9}$$

Following the encoder is the feature extraction module with a Conv-LeakyReLU-Conv structure. In this module, both convolution kernels are set to 3×3 with L channels. The LeakyReLU is employed as the activation function.

$$Feature(\cdot) = conv(LeakyReLU(conv(encoder(\cdot)))) \tag{10}$$

After feature extraction, the encoded feature will feed into a regular res-block consisting of the shuffled multi-spectral channel attention module with one short skip connection. The following is a feature fusion block with L convolution kernels, each with a spatial size of 3×3. This feature fusion block can integrate the sum of features with or without passing the module, and then another point-wise convolution is used to decode the feature into the required HSI.

Our network naturally contains skip-connections due to the Eq(7). Since our TEP-module is based on a residual structure, thus the entire network can be regarded as a simplified Residual-In-Residual structure. As [40] illustrates, these short and long skip connections will allow the low-frequency information to be passed backward through them. The main network will be allowed to concentrate on learning high-frequency information, which is consistent with our intention.

Shuffled Multi-spectral Channel Attention (Shuffled-MCA). We aim to enhance the high-frequency information of the data while completing it. So we try to introduce frequency domain information. We refer to the settings of [20] to set up our shuffled multi-spectral channel attention. The fed feature $\mathcal{X}^K$ is first divided into g groups, i.e., $\mathcal{X}^K = \{\mathcal{X}_0^K, \mathcal{X}_1^K, \cdots, \mathcal{X}_{g-1}^K\}$, each group will be assigned a group of 2D DCT basis generated by one corresponding frequency component. Thus,

$$Freq^i = 2DDCT^{u_i,v_i}(X^i) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} X_{:,h,w}^i B_{h,w}^{u_i,v_i} \tag{11}$$

where $i \in \{0, 1, \cdots, g-1\}$. Then all the $Freq^i$ will be concatenate into one vector, then we got the output of the 'Squeeze' part. By assigning each group with different frequency component, it could effectively weight different information at the same time.

In the previous approaches, the channel descriptors were fed directly into the subsequent 'Excitation' part. However, we assume that although the use of different frequency information is considered, the data within each group still only extracts information under the same frequency component. Therefore, to further improve the 'Squeeze' capability, rather than using channel-wise improvement like[15], we employ the channel shuffle operation and develop the novel *Shuffled Multi-spectral Channel Attention*. Fig.3(a) shows the difference among the structures of our proposed Shuffled-MCA , the original SENet, and FcaNet.

In Fig.3(b), there is a simple illustration of the actual operation of the channel shuffle. The input feature $\mathcal{F}^n$ will be split into g groups, $\{\mathcal{F}_1^1, \mathcal{F}_2^1, \cdots, \mathcal{F}_i^g\}$, where i represents the number of channels within each group. Then, rearranging the order of $\{\mathcal{F}_1^1, \mathcal{F}_2^1, \cdots, \mathcal{F}_i^g\}$, the output of the channel shuffle would be $\{\mathcal{F}_1^1, \mathcal{F}_1^2, \cdots, \mathcal{F}_1^g, \mathcal{F}_2^1, \cdots, \mathcal{F}_i^g\}$.

We added the channel shuffle operation between 'Squeeze' and 'Excitation'. This operation allows the extraction of information corroding to different frequency components not limited within adjacent channels. By rearranging the order of channel descriptors it will affect the whole data in as extensive a range as possible. We also provide ablation experiments about the embedding strategies of channel shuffle. In Fig.3(c), the numerical experiments confirm that the channel shuffle operation significantly improves the model's applicability.

4 Experiments

This section will present the details of experiments. First, section 4.1 introduces the database, experimental setup, etc. Section 4.2 offers the results of numerical experiments and the visualization results. Section 4.3 is the ablation experiments related to TEP-net.

4.1 Network Training

In our experiments, two hyperspectral datasets, CAVE[31] and KAIST[4], are used. We select a portion of the data from CAVE for training, and randomly crop out a patch with a spatial scale of 96×96 as one of the training samples. After modulating, the data will be spatially shifted by two-pixel intervals. Then, the spectral dimensions of the shifted data are summed up to produce a two-dimensional measurement of size 96×150 as one of the network input. In order to augment the dataset, each training sample will be randomly flipped or rotated. Meanwhile, 10 images are selected from the KAIST as the test set. For validation purposes, the KAIST will not be present in the training samples. The KAIST images were cropped to a spatial size of 256×256. Also, instead of using the traditional binary mask, the real mask obtained from measurements in the real

system was used[16]. To be consistent with the real system, we choose 28 out of 31 bands for the experiments, with a spectral range of 450nm to 650nm.

We train the network using an end-to-end training approach, where the training objective is minimizing the Root Mean Square Error (RMSE) between the reconstructed data and the original data. The model is trained by the Adam optimizer with the learning rate 0.001. Given that the network uses ReLU or LeakyReLU as the activation function, the Kaiming Normal[8] was chosen to initialize the convolutional layers. The experiments were implemented by PyTorch and trained with a single Nvidia GeForce RTX 2080Ti. The experiments are set to 200 epochs and take about 20 hours.

4.2 Experiment Comparison

We compare various algorithms, including the traditional iterative algorithm: TwIST[2],GAP-TV[32] , the deep neural networks: TSAnet[16], PnP-DIP[17] and the deep unfolding networks: HSSP[25], DGSMP[10]. For all methods, we use the source code released by the original authors, with TSAnet using their subsequent supplemented PyTorch version, and we rewrite HSSP in PyTorch. All learning-based methods have been retrained using the same training set.

Numerical Results The Table 1(1) shows the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) results for each image on the KAIST dataset. Our advantage over the iterative algorithm is evident, with a lead of 8.65dB to 9.95dB in PSNR and 0.2471 to 0.2488 in SSIM. Moreover, for the learning-based approaches, our method is 6.06dB, 4.69dB higher than HSSP and TSAnet in the average PSNR , and is 0.0371 and 0.0694 higher in the average SSIM. Compared to DGSMP, the state-of-the-art method, which is also a deep unfolding network, the PSNR is 3.27dB higher, and SSIM is 0.037 higher. Due to equipment and time constraints, we only use 30 images from the CAVE dataset to build the training set and use data augmentation to augment the training samples to 5000 per epoch, which may explain the performance degradation of TSAnet and DGSMP. However, our performance also illustrates the proposed method's powerful and robust feature capture capability. DGSMP also claims to work on recovering edges and textures, and the effect shows that introducing frequency information may achieve better quality.

Since we are dealing with a hyperspectral image reconstruction task, we further present the results on four indices, Root Mean Squared Error (RMSE), Erreur Relative Globale Adimensionnelle de Synthèse (ERGAS), Spectral Angle Mapper (SAM), and Universal Image Quality Index(UIQI) in Table 1(2). The SAM describes the level of spectral similarity, which is essential to evaluate the quality of hyperspectral reconstruction tasks, and smaller SAM means better spectral fidelity. Among the seven methods, our method has the smallest SAM value, highlighting our performance on spectral reconstruction. Despite its unsatisfactory indices such as PSNR and SSIM, we also note that the conventional method still has comparable spectral fidelity due to its construction approach from domain knowledge.

Table 1. Numerical experiments on KAIST dataset. Best in bold.

(1) The PSNRs(dB) and SSIMs of seven methods. (In each cell, PSNR is on the left and SSIM is on the right.)

	TwIST[2]	GAP-TV[32]	HSSP[25]	TSAnet[16]	PnP-DIP[17]	DGSMP[10]	Shuffled MCA (ours)
scene01	25.25/0.6665	25.64/0.7451	31.44/0.9207	30.47/0.8616	32.04/0.8646	31.19/0.8715	**33.79/0.9117**
scene02	22.04/0.5541	23.41/0.6052	27.08/0.8664	28.96/0.8071	26.03/0.7136	29.97/0.8457	**33.49/0.9038**
scene03	22.20/0.7415	23.14/0.7639	21.17/0.7937	26.79/0.8553	29.83/0.8259	29.97/0.8682	**33.34/0.9033**
scene04	30.33/0.8134	33.04/0.8620	27.81/0.8576	36.89/0.9241	38.03/0.9268	35.33/0.9216	**40.02/0.9463**
scene05	19.27/0.6411	19.11/0.6684	26.76/0.8902	26.53/0.8261	28.80/0.8438	27.55/0.8357	**30.75/0.8951**
scene06	24.29/0.6529	26.69/0.7430	31.35/**0.9277**	29.21/0.8688	29.43/0.8391	30.54/0.9074	**33.17**/0.9241
scene07	18.26/0.5394	22.10/0.6357	27.97/0.8499	25.68/0.7606	26.84/0.7927	27.86/0.8194	**31.75/0.8870**
scene08	26.13/0.6985	25.23/0.2112	27.60/0.8832	26.57/0.8389	28.24/0.8302	29.51/0.8840	**31.77/0.9015**
scene09	22.18/0.7094	24.30/0.6964	21.58/0.7892	28.23/0.8576	28.78/0.8704	29.86/0.8802	**33.55/0.9137**
scene10	22.60/0.5750	23.95/0.6434	29.73/**0.9387**	26.85/0.7940	27.52/0.7843	28.67/0.8840	**31.46**/0.9015
Average	23.36/0.6592	24.66/0.6575	27.25/0.8717	28.62/0.8394	29.55/0.8291	30.04/0.8718	**33.31/0.9088**

(2) The RMSEs, ERGASs, SAMs and UIQIs of seven methods.

	TwIST[2]	GAP-TV[32]	HSSP[25]	TSAnet[16]	PnP-DIP[17]	DGSMP[10]	Shuffled MCA (ours)
RMSE	27.22	21.43	13.63	10.2	14.24	8.51	**5.85**
ERGAS	119.45	93.54	74.86	50.49	62.87	45.37	**29.88**
SAM	15.54	14.81	20.76	13.4	14.67	13.43	**11.59**
UIQI	0.4247	0.5473	0.6974	0.6648	0.6672	0.7205	**0.7641**

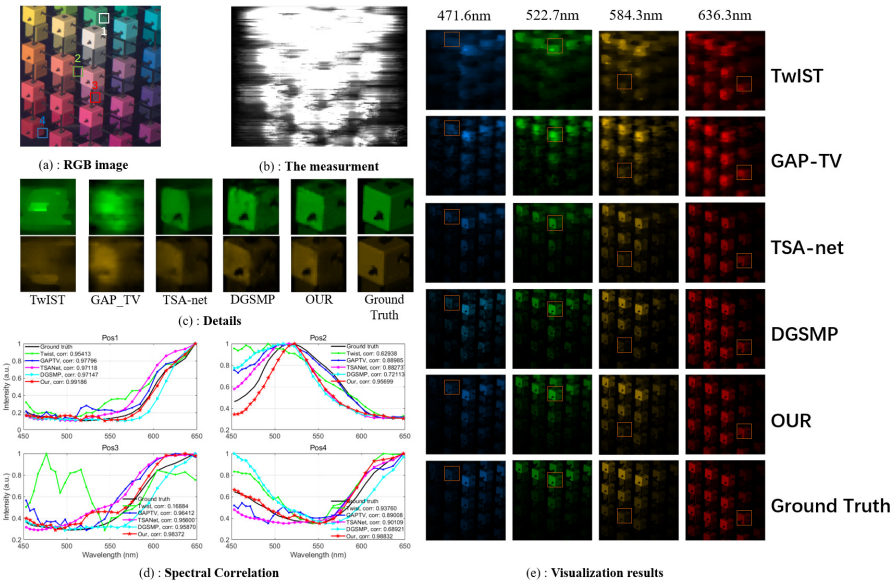


Fig. 4. The visualization results for scene02.(a) The corresponding RGB image. (b) The measurements to be recovered.(c) Zooming in details. (d) Spectral correlation with 4 positions. (e) The visualization results for scene02.

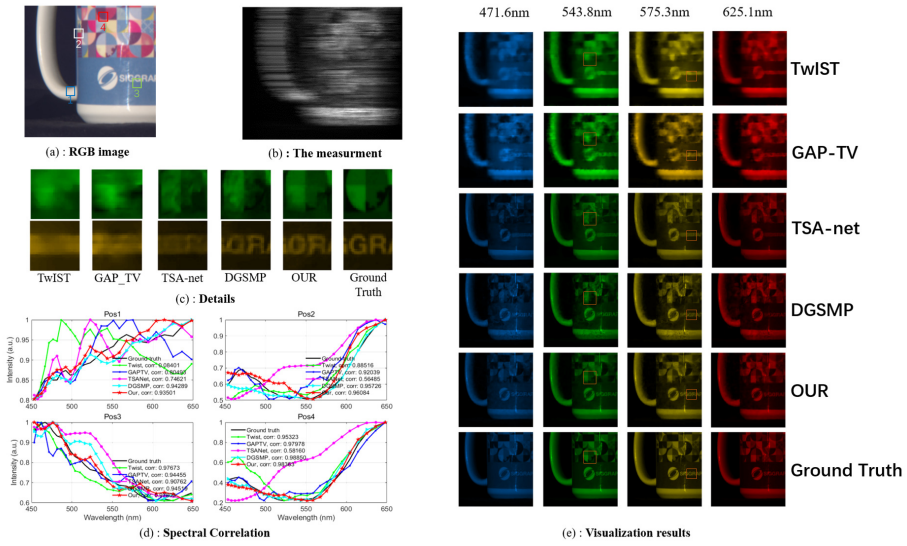
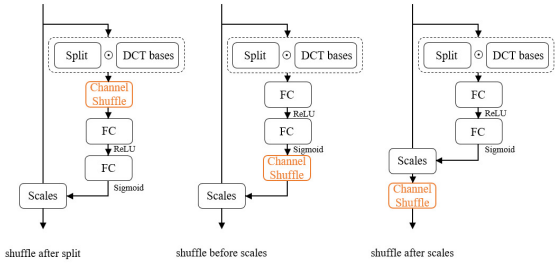


Fig. 5. The visualization results for scene05.(a) The corresponding RGB image. (b) The measurements to be recovered.(c) Zooming in details. (d) Spectral correlation with 4 positions. (e) The visualization results for scene05.

Visualization. TEP-net introduces a residual shuffled multi-spectral channel attention mechanism into the framework of deep unfolding networks and places the hyperspectral image decoding task under the frequency domain. Unlike low-rank theory and CNNs, which tend to deal more with low-frequency information, Shuffle-MCA, by assigning different frequency components to different channels, recovers low-frequency information and enhances high-frequency information synchronously. Therefore, our algorithm can effectively enhance the texture, edges and other regions of the image with significant gradient changes. Fortunately, the visualization supports our hypothesis.

In Fig.4. and Fig.5, the (a) and (b) are the RGB images and the measurements of the testing data, respectively. (d) shows our spectral recovery results, and the numerical indicators are the correlation coefficients of each pair of spectral response curves between the recovered data and the real data. Higher values indicate a higher correlation. (e) shows the visualization results obtained by TwIST[2], GAP-TV[32], TSAnet[16], DGSMP[10] and ours TEP-net, respectively.

For demonstration purposes, we select 4 of the 28 bands. They show that the results of our image recovery are much closer to the real data, and to clarify this we also select patches for better view. We are zooming to show that the results of both iterative methods only have the general shape, with TSA-net being too smooth to lose details. Although DGSMP also claims that they are focusing on edge and texture recovering, their results still shows unpleasant artifacts at the



(a) Three strategies for the embedding.

(b) The embedded position of the channel shuffle.

	After split	Before scales	After scales
PSNR(dB)	**33.31**	32.81	32.43
SSIM	**0.9088**	0.8983	0.8992

(c) The group number of channel shuffle

	g = 4	g = 8	g = 16
PSNR	33.18	33.28	**33.31**
SSIM	0.9103	0.9094	**0.9088**

Fig. 6. Ablation experiments on the embedding of the channel shuffle.

edges and are less stable at complex textures. In contrast, our results recover both edges and appearance well. Texture enhancement does not mean losing the capture of low-frequency information. We assume that the residuals of the sub-network and the self-contained residual structure of the inference process allow low-frequency information to circulate efficiently over the network. It is easy to see that our method still achieves superior performance in smoother areas, such as the bottom of the cup and the side of the cube. More results will show in the supplementary material.

4.3 Ablation Experiments

The Embedding of the Channel Shuffle. In order to solve the imbalance of information extraction due to the grouping, the channel shuffle operation will be added to enhance the network performance. However, it is debatable where to add it, and we finally decided to add it into the middle of 'Squeeze' and 'Excitation'. Hence, it will be more beneficial for information extraction by rearranging channel descriptors obtained by 'Squeeze' and then putting them into 'Excitation' to find the correlation between channels. To verify the effectiveness, we did a set of ablation experiments.

Fig.6. shows the comparison of the three embedding forms. When shuffling after scales, the results dropped, which is because when shuffling after scales, the output feature's channels do not correspond to the residuals anymore, leading to a mixture of non-corresponding information and resulting in an unpleasant consequence. Even though the channel shuffle can always have improvement during the generation of the attention, shuffle before 'Excitation' still got the best performance, which also validates our previous point.

We also offer the selection of hyper-parameters in the channel shuffle , the grouping g. According to Fig.6.(c), the trends of PSNR and SSIM obtained under different strategies are not consistent, so we chose the strategy $g = 16$ by observing the generated images. The specific reference images are in the supplementary material.

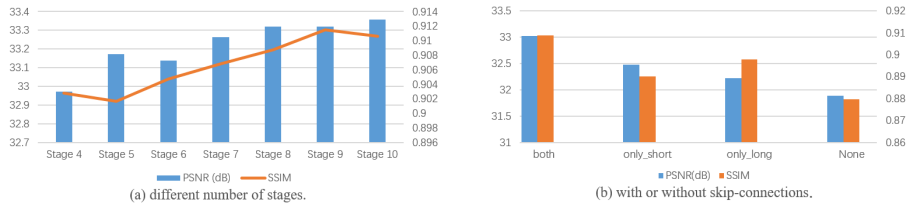


Fig. 7. Reconstruction results of (a) different number of stages. (b) with or without skip-connections.

The Number of Stages. In deep unfolding networks, we combine traditional iterative algorithms with deep learning methods, viewing one iteration as one stage, and connecting K stages in series to construct the network. From this perspective, the selection of K is critical, and the selection of K varies in different tasks. Fig.7(a), a larger K may lead to a better results, while takes a much longer to converge. In order to balance computational cost and effectiveness, $K = 8$ is selected for the experiments in this paper.

The skip-connections. As we stated previously that our scheme embeds the channel attention mechanism into a residual structure in the hope of propagating low frequency information backwards through the residual skipping connection, in combination with the MCA module to obtain better recovery. In Fig.7(b), we show the results with or without the skipping connections. We keep the long or short skipping connections alone, or just the backbone network, and there is a decrease in performance, indicating that the residual structure took effect in our task.

5 Conclusion

In this paper, we propose an HQS-based decoding method and expand its inference into a deep unfolding network that can be trained end-to-end for snapshot compressive imaging. Then we propose a residual shuffled MCA sub-network for texture enhancement prior learning and treat it as a learnable proximal operator of the model. By introducing the DCT bases, the model could effectively enhance the information corresponding to the different frequency components. Meanwhile, to overcome the shortcoming caused by the grouping operation within the MCA, we employ the channel shuffle operation to improve the robustness of the network by rearranging the channel descriptors' order. The experiments demonstrate that we have significantly improved the numerical evaluation metrics compared to the state-of-the-art methods. The visualization results also verify the superiority of our texture enhancement learning effect.

References

1. Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al.: A closer look at memorization in deep networks. In: International conference on machine learning. pp. 233–242. PMLR (2017)
2. Bioucas-Dias, J.M., Figueiredo, M.A.: A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration. IEEE Transactions on Image processing **16**(12), 2992–3004 (2007)
3. Cheng, Z., Chen, B., Liu, G., Zhang, H., Lu, R., Wang, Z., Yuan, X.: Memory-efficient network for large-scale video compressive sensing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16246–16255 (2021)
4. Choi, I., Jeon, D.S., Nam, G., Gutierrez, D., Kim, M.H.: High-quality hyperspectral reconstruction using a spectral prior. ACM Transactions on Graphics (TOG) **36**(6), 1–13 (2017)
5. Gehm, M.E., John, R., Brady, D.J., Willett, R.M., Schulz, T.J.: Single-shot compressive spectral imaging with a dual-disperser architecture. Optics express **15**(21), 14013–14027 (2007)
6. Govender, M., Chetty, K., Bulcock, H.: A review of hyperspectral remote sensing and its application in vegetation and water resource studies. Water Sa **33**(2), 145–151 (2007)
7. Gregor, K., LeCun, Y.: Learning fast approximations of sparse coding. In: Proceedings of the 27th international conference on international conference on machine learning. pp. 399–406 (2010)
8. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. pp. 1026–1034 (2015)
9. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
10. Huang, T., Dong, W., Yuan, X., Wu, J., Shi, G.: Deep gaussian scale mixture prior for spectral compressive imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16216–16225 (2021)
11. Liu, Y., Yuan, X., Suo, J., Brady, D.J., Dai, Q.: Rank minimization for snapshot compressive imaging. IEEE transactions on pattern analysis and machine intelligence **41**(12), 2990–3006 (2018)
12. Lu, B., Dao, P.D., Liu, J., He, Y., Shang, J.: Recent advances of hyperspectral imaging technology and applications in agriculture. Remote Sensing **12**(16), 2659 (2020)
13. Ma, J., Liu, X.Y., Shou, Z., Yuan, X.: Deep tensor admm-net for snapshot compressive imaging. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10223–10232 (2019)
14. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: Proceedings of the European conference on computer vision (ECCV). pp. 116–131 (2018)
15. Magid, S.A., Zhang, Y., Wei, D., Jang, W.D., Lin, Z., Fu, Y., Pfister, H.: Dynamic high-pass filtering and multi-spectral attention for image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4288–4297 (2021)

16. Meng, Z., Ma, J., Yuan, X.: End-to-end low cost compressive spectral imaging with spatial-spectral self-attention. In: European Conference on Computer Vision. pp. 187–204. Springer (2020)
17. Meng, Z., Yu, Z., Xu, K., Yuan, X.: Self-supervised neural networks for spectral snapshot compressive imaging. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2622–2631 (2021)
18. Miao, X., Yuan, X., Pu, Y., Athitsos, V.: l-net: Reconstruct hyperspectral images from a snapshot measurement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4059–4069 (2019)
19. Park, J., Woo, S., Lee, J.Y., Kweon, I.S.: Bam: Bottleneck attention module. arXiv preprint arXiv:1807.06514 (2018)
20. Qin, Z., Zhang, P., Wu, F., Li, X.: Fcanet: Frequency channel attention networks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 783–792 (2021)
21. Su, K., Yu, D., Xu, Z., Geng, X., Wang, C.: Multi-person pose estimation with enhanced channel-wise and spatial information. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5674–5682 (2019)
22. Wagadarikar, A., John, R., Willett, R., Brady, D.: Single disperser design for coded aperture snapshot spectral imaging. Applied optics **47**(10), B44–B51 (2008)
23. Wagadarikar, A.A., Pitsianis, N.P., Sun, X., Brady, D.J.: Video rate spectral imaging using a coded aperture snapshot spectral imager. Optics express **17**(8), 6368–6388 (2009)
24. Wang, H., Wu, X., Huang, Z., Xing, E.P.: High-frequency component helps explain the generalization of convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8684–8694 (2020)
25. Wang, L., Sun, C., Fu, Y., Kim, M.H., Huang, H.: Hyperspectral image reconstruction using a deep spatial-spectral prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8032–8041 (2019)
26. Wang, X., Kan, M., Shan, S., Chen, X.: Fully learnable group convolution for acceleration of deep neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9049–9058 (2019)
27. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
28. Xiong, Z., Shi, Z., Li, H., Wang, L., Liu, D., Wu, F.: Hscnn: Cnn-based hyperspectral image recovery from spectrally undersampled projections. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 518–525 (2017)
29. Yang, J., Yuan, X., Liao, X., Llull, P., Sapiro, G., Brady, D.J., Carin, L.: Gaussian mixture model for video compressive sensing. In: 2013 IEEE International Conference on Image Processing. pp. 19–23. IEEE (2013)
30. Yang, Y., Sun, J., Li, H., Xu, Z.: Deep admm-net for compressive sensing mri. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. pp. 10–18 (2016)
31. Yasuma, F., Mitsunaga, T., Iso, D., Nayar, S.K.: Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum. IEEE transactions on image processing **19**(9), 2241–2253 (2010)
32. Yuan, X.: Generalized alternating projection based total variation minimization for compressive sensing. In: 2016 IEEE International Conference on Image Processing (ICIP). pp. 2539–2543. IEEE (2016)

33. Yuen, P.W., Richardson, M.: An introduction to hyperspectral imaging and its application for security, surveillance and target acquisition. The Imaging Science Journal **58**(5), 241–253 (2010)
34. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning (still) requires rethinking generalization. Communications of the ACM **64**(3), 107–115 (2021)
35. Zhang, J., Zhao, H., Yao, A., Chen, Y., Zhang, L., Liao, H.: Efficient semantic scene completion network with spatial group convolution. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 733–749 (2018)
36. Zhang, J., Ghanem, B.: Ista-net: Interpretable optimization-inspired deep network for image compressive sensing. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1828–1837 (2018)
37. Zhang, K., Gool, L.V., Timofte, R.: Deep unfolding network for image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3217–3226 (2020)
38. Zhang, Q.L., Yang, Y.B.: Sa-net: Shuffle attention for deep convolutional neural networks. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2235–2239. IEEE (2021)
39. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6848–6856 (2018)
40. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV). pp. 286–301 (2018)