Hyperspectral Image Super-resolution via Knowledge-Driven Deep Unrolling and Transformer Embedded Convolutional Recurrent Neural Network

Kaidong Wang, Xiuwu Liao, Jun Li, *Fellow, IEEE*, Deyu Meng, *Member, IEEE*, and Yao Wang

***Abstract*—Hyperspectral (HS) imaging has been widely used in various real application problems. However, due to the hardware limitations, the obtained HS images usually have low spatial resolution, which could obviously degrade their performance. Through fusing a low spatial resolution HS image with a high spatial resolution auxiliary image (e.g., multispectral, RGB or panchromatic image), the so-called HS image fusion has underpinned much of recent progress in enhancing the spatial resolution of HS image. Nonetheless, a corresponding well registered auxiliary image cannot always be available in some real situations. To remedy this issue, we propose in this paper a newly single HS image super-resolution method based on a novel knowledge-driven deep unrolling technique. Precisely, we first propose a maximum a posterior based energy model with implicit priors, which can be solved by alternating optimization to determine an elementary iteration mechanism. We then unroll such iteration mechanism with an ingenious Transformer embedded convolutional recurrent neural network in which two structural designs are integrated. That is, the vision Transformer and 3D convolution learn the implicit spatial-spectral priors, and the recurrent hidden connections over iterations model the recurrence of the iterative reconstruction stages. Thus, an effective knowledge-driven, end-to-end and data-dependent HS image super-resolution framework can be successfully attained. Extensive experiments on three HS image datasets demonstrate the superiority of the proposed method over several state-of-the-art HS image super-resolution methods.**

***Index Terms*—Hyperspectral (HS) image, super-resolution (SR), deep unrolling, convolutional recurrent neural network (CRNN), spatial-spectral priors.**

I. Introduction

HYPERSPECTRAL sensors can simultaneously acquire images of one scene in tens to hundreds of contiguous and narrow spectral bands of the electromagnetic spectrum. As a consequence, compared with the natural image which has only three bands with red, green and blue, or the multispectral (MS) image which has several bands, the HS image provides abundant and detailed spectral information regarding the physical nature of different materials presented in the scene. Benefiting from this property, HS images have been extensively exploited in a wide spread of relevant applications, including object detection [1], military [2], tracking [3], land surface classification [4] and monitoring [5]. Unfortunately, due to the hardware limitations, the very high spectral resolution of HS image is at the expense of the severe limitations in its spatial resolution. Actually, because of the dense spectral bands in the HS imaging systems, the average amount of photons reached a narrow spectral window is quite limited, and thus lots of exposures are usually necessary to ensure sufficient signal-to-noise ratio (SNR), which usually leads to the sacrifice of spatial resolution [6]. The resulting poor spatial resolution may degrade the performances to some extent in a number of applications. To alleviate this issue, the task of spatial resolution enhancement for HS imaging has received much attention and several related studies have been conducted in the past few years [7]. It is worth mentioning that as a straightforward way, simply increasing the spatial resolution of HS sensors usually cannot hit the mark. Therefore, aiming to convert a low spatial resolution HS image to a high resolution one by postprocessing, HS image super-resolution is a software manner without modifying any hardware and thus provides a better alternative to the former.

On the basis of the different input sources, there are mainly two kinds of methods to implement the spatial enhancement of HS images, i.e., fusion based HS image super-resolution method and single HS image super-resolution method. The former is to improve the spatial resolution of a low spatial resolution HS image by merging it with a high spatial resolution and low spectral resolution auxiliary image of the same scene [8]. The panchromatic image is a class of frequently-used auxiliary image which is usually fused with the corresponding low resolution MS image to enhance its resolution, and this procedure is called pan-sharpening [9]. Conventional pan-sharpening methods can be roughly divided into three categories, i.e., sparse representation (SR) [10], component substitution (CS) [11] and multiresolution analysis (MRA) [12]. Except for the conventional model based approaches, deep learning based methods have also been introduced to pan-sharpening in recent years and achieved impressive performance [13], [14]. Although achieving higher spatial resolution HS images, the results produced by pan-sharpening methods often suffer from serious spectral distortion due to the lackness of spectral information in the corresponding panchromatic images. To address this problem, auxiliary images with more spectral information are necessary.

K. Wang, X. Liao and Y. Wang are with the Center for Intelligent Decision-Making and Machine Learning, School of Management, Xi'an Jiaotong University, Xi'an 710049, China (e-mails: wangkd13@gmail.com; yao.s.wang@gmail.com; liaoxiuwu@xjtu.edu.cn).

J. Li is with the Guangdong Provincial Key Laboratory of Urbanization and Geo-Simulation, School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China (e-mail: lijun48@mail.sysu.edu.cn).

D. Meng is with the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China, also with the Key Laboratory of Intelligent Networks and Network Security of Ministry of Education, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: dymeng@xjtu.edu.cn).

Compared with the panchromatic image which records wavelengths in only one band, the MS image with several bands contains more spectral information, and thus usually provides a better alternative. This so-called HS and MS image fusion is to fuse the low resolution HS image with a high resolution MS image of the same scene to enhance the spatial resolution of the former. Different from the aforementioned pan-sharpening technique, the high spatial resolution HS image generated by HS and MS image fusion can often enhance the image detail while maintaining excellent spectral information. As such, a great number of interesting works have devoted to the task of HS and MS image fusion. See, e.g., [15], [16], [17], [18], [19], [20], [21], [22], among others.

Despite the proven efficacy of fusion based HS image super-resolution method, a well registered high spatial resolution auxiliary image is necessary for this kind of method. However, such an assumption is difficult to conform with in practice [23]. Hence, it is worthwhile exploring to directly improve the spatial resolution of the input low resolution HS image without using any auxiliary information. This line of research is often termed as single HS image super-resolution. In general, the existing single HS image super-resolution methods can be roughly divided into two categories, i.e., knowledge-driven model optimization and data-driven deep learning approaches. The former usually considers the low resolution HS image as the prior knowledge of the high resolution one, and thus builds the reconstruction model by using some regularizers to exploit the prior knowledge. For example, [24] considered the Total Variation (TV) regularization term and [25] used a spatial sparsity term and a local spectral similarity preserving term as the regularization term.

In the past decade, due to the advent of the large-scale image databases and the inexorable growth of computing power, deep learning based approaches have shown great superiority to the conventional knowledge-driven methods in many computer vision tasks [26], [27], especially image super-resolution task [28], [29]. Most of the existing deep learning based methods are originally designed for conventional gray-scale or RGB images, e.g., the super-resolution convolutional neural network (SRCNN) [30], the super-resolution generative adversarial network (SRGAN) [31], deep recursive residual network (DRRN) [32], among numerous others. Basically, these deep learning based methods can be directly employed to enhance the spatial resolution of HS images in a band-by-band manner. However, due to being failed to exploit the inherent structure, especially the spectral correlation, underlying the HS images, those extensions often suffer from severe spectral distortion, and thus degrade the reconstruction performance. To address this problem, a lot of efforts have been made in the recent years, leading to many new approaches aiming specially at the single HS image super-resolution task [33], [34], [35].

The aforementioned two categories of single HS image super-resolution methods, i.e., knowledge-driven model optimization and data-driven deep learning approaches, have their own intrinsic limitations. For the former, some reasonable prior knowledge about the targeted high resolution HS image is necessary, and the fit between the prior knowledge and the groundtruth directly determines the reconstruction performance. Unfortunately, such prior knowledge accurately describing the inherent structure of the targeted HS image is not easy to exploit in a concise mathematical language. For the latter, the deep structure is usually designed heuristically as a mapping between the low resolution HS image and the corresponding high resolution one. This procedure could abandon the domain knowledge, resulting in the lackness of consideration of problem characteristics.

To inherit the advantages and meanwhile avoid the defects of the both methodologies, we shall combine the knowledge-driven model optimization and the data-driven deep learning methods and propose a novel knowledge-driven deep unrolling framework for the single HS image super-resolution task. To avoid the perplexing prior knowledge exploiting, we consider a maximum a posterior (MAP) based energy model with implicit priors to be learned from training data. Experiments show that the learning-based priors are capable of describing the characteristics of the targeted data more accurately compared to the conventional artificially designed ones. Then an alternating optimization based iterative algorithm is developed to solve the resulting model, giving rise to an elementary iteration mechanism. This iteration mechanism is then unrolled with an ingenious Transformer embedded convolutional recurrent neural network which is a combination of two structures: (1) vision Transformer and 3D convolution are employed to learn the implicit spatial-spectral priors; (2) recurrent hidden connections over iterations are exploited to model the recurrence of the iterative reconstruction stages. As a consequence, we get an effective knowledge-driven, end-to-end and data-dependent HS image super-resolution framework, in which the entire feed-forward structure is actually guided by the model rather than heuristically designed as done in other deep learning based methods.

The rest of this paper is organized as follows. Section II presents the related work. The proposed knowledge-driven deep unrolling framework is introduced in Section III. Extensive experimental results on three image datasets are presented in Section IV. And some concluding remarks are drawn in Section V.

II. Related Work

In this section, we briefly review some existing methods most relevant to our work, including model optimization and deep learning based HS image super-resolution method, deep unrolling strategy and vision Transformer.

A. Model Optimization based HS image super-resolution

Model optimization based methods for HS image super-resolution usually consider the low resolution image as the prior knowledge of the low resolution one, and regularize the solution space using prior assumptions. An early representative work is [36] in which the hyperspectral observations from different wavelengths are represented as weighted linear combinations of a small number of basis image planes, and then the spectrum of the observed scene is reconstructed by fusing information from multiple observations and spectral bands. In [37], Huang *et al.* propose a super-resolution approach

with unknown blurring by imposing the low-rank model with predefined spectral subspace and group sparse model to utilize the shared spatial structure across all spectral bands. Xu *et al.* [38] employs a joint spectral-spatial sub-pixel mapping model to obtain the probabilities of sub-pixels belonging to different land cover classes and further generate the resolution enhanced image. He *et al.* [39] employ the tensor nuclear norm and 3D total variation to characterize the global spatial-and-spectral correlation and local smoothness of the targeted image, respectively. In [40], Wang *et al.* propose a tensor based approach to exploit the three intrinsic characteristics of HS image, i.e., the global correlation across spectral domain, the nonlocal self-similarity across spatial domain, and the local smooth structure across both spatial and spectral domains, and achieve superior reconstruction performance. A MAP based approach is proposed in [41] which converts the ill-posed reconstruction problem in the spectral domain to a quadratic optimization problem in the abundance map domain. In addition, sparse representations and dictionary learning based approaches are widely employed in this field [37], [42].

B. *Deep Learning based HS image supe-rresolution*

Deep learning based approaches for HS image super-resolution have been widely studied in the recent few years. In [43], deep convolutional neural network is firstly introduced into single HS image super-resolution problem, which transfer the learned mapping between low and high resolution images from RGB image domain to HS image domain. Taking that HS image contains abundant and detailed spectral information in consideration, Mei *et al.* [33] first introduce 3D convolution to simultaneously exploit the spatial context of neighboring pixels and spectral correlation of neighboring bands in HS images, resulting in a three-dimensional full convolutional neural network (3DFCNN). The regular 3D convolution used in [33] leads to a significant increase in network parameters. To address this issue, Wang *et al.* [44] propose a spectral-spatial residual network (SSRNet) employing spatial and temporal separable 3D convolution to effectively explore spatial-spectral information and meanwhile reduce unaffordable memory usage. Li *et al.* [45] propose a mixed convolutional network (MCNet) to extract the potential features by 2D/3D convolution instead of one convolution. A similar thought exploring the relationship between 2D/3D convolution (ERCSR) is proposed in [34]. Inspired by the high similarity among adjacent bands, in [46] Wang *et al.* designs a dual-channel network through 2D and 3D convolution to jointly exploit the information from both single band and adjacent bands. To fully exploit the spatial and spectral prior of HS image, Jiang *et al.* [47] design a spatial-spectral block consisting of a spatial residual module and a spectral attention residual module, and further propose a group convolution and progressive upsampling framework for the stable network training.

C. *Deep Unrolling Strategy*

Deep unrolling strategy has attracted widespread attention in the last few years owing to its ability to provide a concrete and systematic connection between iterative algorithms and deep neural networks[48]. A typical deep unrolling process includes building the optimization model, constructing an iterative solving algorithm, embedding learnable modules (usually neural networks) in this iterative algorithm, and finally developing an potentially efficient, high-performance and yet interpretable network architecture [49], [50]. LISTA [51] is the first deep unrolling based framework which unfolds the iterative shrinkage thresholding algorithm to a non-linear, feed-forward predictor for sparse coding. From then on, deep unrolling strategy has been flourishing and exploited in many applications, including background foreground separation [49], video reconstruction [50] and low-light image enhancement [52]. Meanwhile, there have been some relevant studies in the HS image reconstruction field, especially the multispectral and hyperspectral image fusion (MS/HS fusion) task[53], [54]. In [53], Xie *et al.* construct an MS/HS fusion model merging the generalization models of low resolution images and the low-rankness prior knowledge into a concise formulation, and then unfold the proximal gradient algorithm for solving the proposed model to design the network architecture. Dong *et al.* [54] propose an iterative MS/HS fusion algorithm based on a deep HS image denoiser to leverage both domain knowledge likelihood and deep image prior. Except for the MS/HS fusion, Huang *et al.* [55] propose an Maximum a Posterior (MAP) estimation framework using learned Gaussian Scale Mixture (GSM) prior for the coded aperture snapshot spectral imaging (CASSI) reconstruction task, and Ma *et al.* [56] unfold the iterative process of an alternative direction multiplier method (ADMM) algorithm into a multistage network for the spatiospectral image super-resolution (SSSR) task, i.e., joint spatial and spectral super-resolution. As far as we know, there has been no deep unrolling based approach aiming specially at the single HS image super-resolution task, although [56] can be transferred to handle this problem.

D. *Vision Transformer*

Transformer is originally proposed by Vaswani *et al.* [57] for natural language processing (NLP). Due to its impressive performance, Transformer is further used in vision problems as an alternative to CNN, including objection detection [58], segmentation [59] and image classification [60]. In the vision problem, Transformer learns to attend to important image regions by exploring the global interactions between different regions, and thus shows promising performance. Chen *et al.* [61] propose a backbone model IPT for image restoration based on the standard Transformer. However, IPT relies on a huge amount of parameters and calculations, and thus is difficult to train and applied in large size images. To address this, Liu *et al.* [62] proposed a hierarchical Transformer named Swin Transformer whose representation is computed with Shifted windows. Owing to its shifted windowing scheme, Swin Transformer can both model long-range dependency like standard Transformer and process high resolution images like CNN. Following [62], Liang *et al.* [63] propose SwinIR for image restoration based on the Swin Transformer, which achieves state-of-the-art performance in various image reconstruction tasks.

III. Proposed Framework

Throughout this paper, the high resolution HS image to be reconstructed is denoted as a third-order tensor $\mathcal{X} \in \mathbb{R}^{W\times H\times S}$ for the subsequent unrolling procedure, where W, H and S are the dimensions of the width, height and spectral mode, respectively. Correspondingly, the low resolution HS image is denoted as $\mathcal{Y} \in \mathbb{R}^{w\times h\times S}$, where $w = \frac{1}{l_1}W$, $h = \frac{1}{l_2}H$, l_1 and l_2 are the down-sampling factors in horizontal and vertical directions, respectively. As for the HS image super-resolution problem, the low resolution image $\mathcal{Y}$ is usually considered as the spatial degradation of the high resolution one, then in tensor form it can be written as [15], [16]

$$\mathcal{Y} = \mathcal{X} \times_1 \mathbf{P_1} \times_2 \mathbf{P_2}, \tag{1}$$

where $\mathbf{P_1} \in \mathbb{R}^{w\times W}$ and $\mathbf{P_2} \in \mathbb{R}^{h\times H}$ are the blurring and downsampling matrices along the width and height modes, respectively, and $\times_n, n = 1, 2$ denotes the mode-n product of a tensor by a matrix. Then we would like to reconstruct the $\mathcal{X}$ as precisely as possible based on (1) with $\mathbf{P_1}$ and $\mathbf{P_2}$ unknown. Considering that $\mathcal{X}$ and $\mathcal{Y}$ can be unfolded along the third mode as $\mathbf{X}_{(3)} \in \mathbb{R}^{S\times(W\times H)}$ and $\mathbf{Y}_{(3)} \in \mathbb{R}^{S\times(w\times h)}$, thus $\mathbf{Y}_{(3)}$ can then be expressed as

$$\mathbf{Y}_{(3)} = \mathbf{X}_{(3)}\mathbf{A}, \tag{2}$$

where $\mathbf{A} \in \mathbb{R}^{(W\times H)\times(w\times h)}$ denotes the degradation operator of blurring and downsampling corresponding to the $\mathbf{P_1}$ and $\mathbf{P_2}$. Noting that (1) implicitly demands that matrix $\mathbf{A}$ can be decoupled with respect to the two spatial modes of $\mathcal{X}$, that is,

$$\mathbf{A} = (\mathbf{P_2} \otimes \mathbf{P_1})^T. \tag{3}$$

While such a separability assumption is not necessary in our method because that the spatial degradation operation, including blurring and downsampling, is uniformly expressed by a downsampling network consisting of 2D channel-wise convolutions and average pooling operators.

A. Model and Elementary Iteration Mechanism

Reconstructing the high resolution HS image $\mathcal{X}$ from its degraded version $\mathcal{Y}$ is obviously a highly ill-posed inverse problem, we thus make it well-posed by maximizing the posterior probability of $\mathcal{X}$ based on MAP as follows:

$$\mathcal{X} = \arg\max_{\mathcal{X}} P(\mathcal{X}|\mathcal{Y}) \propto P(\mathcal{Y}|\mathcal{X}) \cdot P(\mathcal{X}), \tag{4}$$

where $P(\mathcal{Y}|\mathcal{X})$ is the conditional probability of the degraded version $\mathcal{Y}$, and $P(\mathcal{X})$ denotes the prior probability of $\mathcal{X}$. By performing a negative logarithmic transformation, (4) has the following equivalent form:

$$\mathcal{X} = \arg\min_{\mathcal{X}} -\log P(\mathcal{Y}|\mathcal{X}) - \log P(\mathcal{X}), \tag{5}$$

which can then be further reformulated as the energy function minimization problem below:

$$\mathcal{X} = \arg\min_{\mathcal{X}} \frac{1}{2}\|\mathcal{Y} - \mathcal{X} \times_1 \mathbf{P_1} \times_2 \mathbf{P_2}\|_F^2 + \Phi(\mathcal{X}), \tag{6}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and $\Phi(\mathcal{X})$ represents the implicit regularization of $\mathcal{X}$. We can see from (6) that the first item represents the fidelity of the model to ensure the data consistency, and meanwhile the second one stands for the implicit regularization term which describes the prior knowledge of $\mathcal{X}$. Through reducing the solution space, the regularization term plays a key role in obtaining a satisfactory solution. There are many frequently-used regularization forms describing various inherent structures of the targeted data, such as the l_1 norm for the sparsity, the total variation (TV) for the smoothness, the nuclear norm for the correlation, etc. Although achieving decent performance, those artificially designed regularization forms are relatively rough and thus cannot fully exploit the prior knowledge. To address this problem, an implicit regularization term $\Phi(\mathcal{X})$ is employed in (6) to learn more accurate prior knowledge of $\mathcal{X}$ from the training data, which will be achieved through the following deep unrolling framework.

The optimization problem (6) can be effectively solved by using the half quadratic splitting (HQS) technique [64] which apply a variable splitting technique to decouple the fidelity term and the regularization term. Introducing an auxiliary variable $\mathcal{Z}$ constrained to be equal to $\mathcal{X}$, (6) can be rewritten as

$$\min_{\mathcal{X},\mathcal{Z}} \frac{1}{2}\|\mathcal{Y} - \mathcal{Z} \times_1 \mathbf{P_1} \times_2 \mathbf{P_2}\|_F^2 + \Phi(\mathcal{X}) + \frac{\eta}{2}\|\mathcal{Z} - \mathcal{X}\|_F^2, \tag{7}$$

where η is the penalty parameter. Thus, it can be alternatively solved via the following iterative scheme:

$$\mathcal{Z}^{(k)} = \arg\min_{\mathcal{Z}} \frac{1}{2}\|\mathcal{Y} - \mathcal{Z} \times_1 \mathbf{P_1} \times_2 \mathbf{P_2}\|_F^2 + \frac{\eta}{2}\|\mathcal{Z} - \mathcal{X}^{(k-1)}\|_F^2, \tag{8}$$

$$\mathcal{X}^{(k)} = \arg\min_{\mathcal{X}} \frac{\eta}{2}\|\mathcal{Z}^{(k)} - \mathcal{X}\|_F^2 + \Phi(\mathcal{X}), \tag{9}$$

where $\mathcal{Z}^{(k)}$ and $\mathcal{X}^{(k)}$ denotes the updated $\mathcal{Z}$ and $\mathcal{X}$ in the k-th iteration step, respectively. Equations (8) and (9) make up the elementary iteration mechanism which can then be unrolled with an ingenious deep neural network to achieve effective implicit prior knowledge learning.

B. Fidelity Layer

In our deep unrolling framework, (8) is considered as a layer with $\mathcal{X}^{(k-1)}$ as its input and $\mathcal{Z}^{(k)}$ as its output. This layer is called fidelity layer (FL) as the role it plays in the final deep neural network. To ensure end-to-end training, the fidelity layer should be able to do the back propagation of gradients. Unfortunately, different from the case where the equation has an analytical solution [65], [66], the back propagation of gradients is not an easy task in our fidelity layer due to the unknown $\mathbf{P_1}$ and $\mathbf{P_2}$. To figure out this problem, we solve (8) using a gradient decent based method, and then unfold it into a network-like structure where the unknown $\mathbf{P_1}$ and $\mathbf{P_2}$ are treated as the learning parameters.

Precisely, in the k-th iteration step, we need to solve the $\mathcal{Z}^k$ with the obtained $\mathcal{X}^{(k-1)}$. Taking $\mathcal{X}^{(k-1)}$ as the initial point, we do a one-step gradient decent operation to update $\mathcal{Z}$ as

$$\mathcal{Z}^{(k)} = \mathcal{X}^{(k-1)} - \rho\nabla f(\mathcal{X}^{(k-1)}), \tag{10}$$

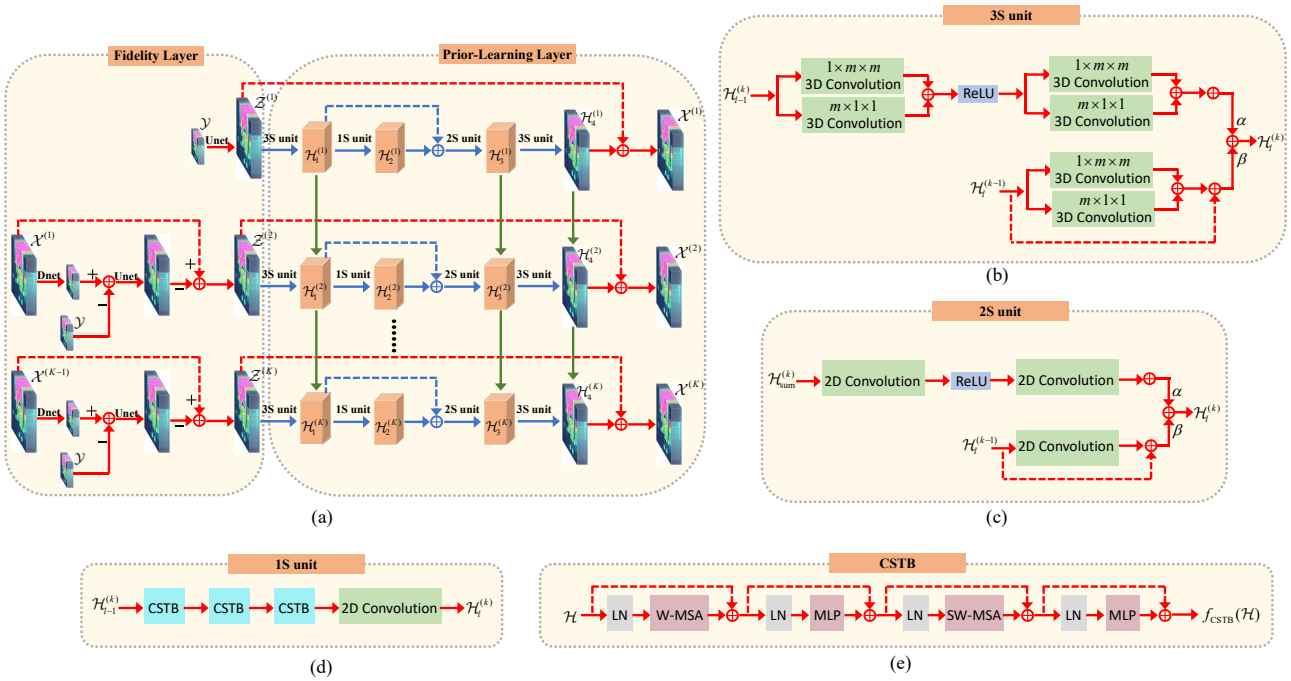


Fig. 1. (a) Flowchart of the proposed deep unrolling based HS image super-resolution framework; (b)-(d) The structure of the 3S, 2S and 1S unit in our framework; (e) The structure of Consecutive Swin Transformer Block (CSTB) in 1S unit.

where ρ is the step size, and $f(\cdot)$ is a function with the following form:

$$f(\mathcal{Z}) = \frac{1}{2}\|\mathcal{Y} - \mathcal{Z} \times_1 \mathbf{P_1} \times_2 \mathbf{P_2}\|_F^2 + \frac{\eta}{2}\|\mathcal{Z} - \mathcal{X}^{(k-1)}\|_F^2. \quad (11)$$

It is easy to see that the gradient of $f(\cdot)$ is

$$\nabla f(\mathcal{Z}) = (\mathcal{Z}\times_1\mathbf{P_1}\times_2\mathbf{P_2} - \mathcal{Y})\times_1\mathbf{P_1}^T\times_2\mathbf{P_2}^T + \eta(\mathcal{Z} - \mathcal{X}^{(k-1)}), \quad (12)$$

where $(\cdot)^T$ denotes the transpose of the matrix. Plugging equation (12) into equation (10), we can easily get the following updating formula of $\mathcal{Z}$:

$$\mathcal{Z}^{(k)} = \mathcal{X}^{(k-1)} - \rho(\mathcal{X}^{(k-1)} \times_1 \mathbf{P_1} \times_2 \mathbf{P_2} - \mathcal{Y}) \times_1 \mathbf{P_1}^T \times_2 \mathbf{P_2}^T. \quad (13)$$

To design our fidelity layer associated with the updating formula of $\mathcal{Z}$, we first decompose (13) into the following four sequential parts:

$$\mathcal{U}^{(k)} = \mathcal{X}^{(k-1)} \times_1 \mathbf{P_1} \times_2 \mathbf{P_2}, \quad (14)$$

$$\mathcal{E}^{(k)} = \mathcal{U}^{(k)} - \mathcal{Y}, \quad (15)$$

$$\mathcal{D}^{(k)} = \rho\mathcal{E}^{(k)} \times_1 \mathbf{P_1}^T \times_2 \mathbf{P_2}^T, \quad (16)$$

$$\mathcal{Z}^{(k)} = \mathcal{X}^{(k-1)} - \mathcal{D}^{(k)}, \quad (17)$$

and then we can design a network to approximately perform the above operations. In equation (14), the mode-1 produce of $\mathcal{X}^{(k-1)}$ by matrix $\mathbf{P_1}$ and the mode-2 produce by $\mathbf{P_2}$ jointly denote the degradation operator of blurring and downsampling on $\mathcal{X}^{(k-1)}$, which can be equivalently represented by 2D convolution and pooling operators in network. And thus it can be performed in the fidelity layer by

$$\mathcal{U}^{(k)} = \mathbf{Dnet}_{\theta_d}(\mathcal{X}^{(k-1)}), \quad (18)$$

where $\mathcal{U}^{(k)} \in \mathbb{R}^{w\times h\times S}$ is the degraded version of $\mathcal{X}^{(k-1)}$ after blurring and downsampling, $\mathbf{Dnet}_{\theta_d}(\cdot)$ denotes the degradation network with θ_d as the learning parameters, which are composed of two parts, i.e., 2D channel-wise convolutions corresponding to the blurring operator, and average pooling corresponding to the downsampling operator. In equations (15) and (17), there are only tensor subtractions which can be calculated directly in the network. As for (16), the mode-1 produce of $\mathcal{E}^{(k)}$ by matrix $\mathbf{P_1}^T$ and the mode-2 produce by $\mathbf{P_2}^T$ jointly represent a spatial upsampling operator which transforms the spatially smaller tensor $\mathcal{E}^{(k)} \in \mathbb{R}^{w\times h\times S}$ to a spatially bigger one $\mathcal{D}^{(k)} \in \mathbb{R}^{W\times H\times S}$. This can be done by the 2D transposed convolution which is the transposition of the combination of convolution and downsampling operator, resulting in the corresponding operation in our fidelity layer below:

$$\mathcal{D}^{(k)} = \mathbf{Unet}_{\theta_u}(\mathcal{E}^{(k)}), \quad (19)$$

where $\mathbf{Unet}_{\theta_u}(\cdot)$ is the spatial upsampling network composed of 2D transposed convolutions, θ_u is the learning parameters, and the ρ in (16) here is coupled into θ_u. Therefore, the equations (18), (15), (19) and (17) are sequentially connected to make up the basic structure of our fidelity layer. In this way, the fidelity layer is able to effectively solve (8), and meanwhile successfully perform the back propagation of gradients.

C. Prior-Learning Layer

After obtaining $\mathcal{Z}^{(k)}$ by solving (8), our next task is to solve the problem (9) which is actually the proximal operator of the prior Φ. As mentioned before, instead of predetermining the explicit form of the regularization term $\Phi(\mathcal{X})$, we shall directly learn the prior knowledge from training data through an ingenious deep neural network to exploit the characteristics

of $\mathcal{X}$. And we use the prior-learning layer (PL) to denote the resulting network. Then it is easy to see that PL should be a deep neural network with $\mathcal{Z}^{(k)}$ as its input and $\mathcal{X}^{(k)}$ as its output. Compared with directly learning a mapping from $\mathcal{Z}^{(k)}$ to $\mathcal{X}^{(k)}$, residual learning is a more effective way in dealing with the image reconstruction problems [67], [68]. With that, we utilize a residual network (ResNet) as the basic form of the prior-learning layer, that is,

$$\mathcal{X}^{(k)} = \mathcal{Z}^{(k)} + \mathbf{Pnet}_{\theta_p}(\mathcal{Z}^{(k)}), \tag{20}$$

where $\mathbf{Pnet}_{\theta_p}(\cdot)$ is the residual learning network with the learning parameters θ_p.

It is obvious that the architecture of $\mathbf{Pnet}_{\theta_p}(\cdot)$ plays an important role in fully exploiting the inherent structure of $\mathcal{X}$, and thus should be carefully designed. After having an ingenious $\mathbf{Pnet}_{\theta_p}(\cdot)$, we can get our knowledge-driven deep unrolling framework established as follows:

$$\begin{aligned}\mathcal{X}^{(0)} \xrightarrow{FL^{(1)}} \mathcal{Z}^{(1)} \xrightarrow{PL^{(1)}} \mathcal{X}^{(1)} \xrightarrow{FL^{(2)}} \mathcal{Z}^{(2)} \xrightarrow{PL^{(2)}} \mathcal{X}^{(2)} \cdots \\ \xrightarrow{PL^{(K-1)}} \mathcal{X}^{(K-1)} \xrightarrow{FL^{(K)}} \mathcal{Z}^{(K)} \xrightarrow{PL^{(K)}} \mathcal{X}^{(K)},\end{aligned} \tag{21}$$

where $FL^{(i)}$ denotes the i-th fidelity layer, $PL^{(i)}$ denotes the i-th prior-learning layer, and procedure (21) actually performs the iteration mechanism in equations (8) and (9) with totally K iteration steps. Hereinafter, an iteration step in (21) is also called a reconstruction stage with the input as the obtained $\mathcal{X}$ in the last stage and the output as the updated $\mathcal{X}$ after the fidelity and prior-learning layer in the current stage.

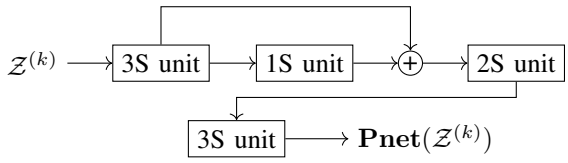
Convolutional neural networks (CNNs) are usually used to learn the spatial information of the traditional images (e.g., gray-scale images and RGB images), and meanwhile vision Transformer has proven to be a promising alternative to CNNs due to its impressive advantage in modeling long-range dependency. In our framework we synthetically utilize vision Transformer and 2D convolution to achieve effective spatial feature extraction of HS image. HS image contains abundant spectral information, and exploiting its spectral correction can obviously help the image reconstruction and reduce the spectral distortion. In our framework we employ 3D convolution to learn the spectral information.

Except for the spatial and spectral information of the HS image $\mathcal{X}^{(k)}$ reconstructed in each stage (k) of (21), there is a third category of contextual information can be learned for a better reconstruction, i.e., the recurrence of the iterative reconstruction stages. Actually, corresponding to an iterative optimization algorithm, the reconstruction process of $\mathcal{X}^{(k)}$ in each stage can be seen as a continuous sequence with stage (k) increasing. Characterizing the continuity of the sequence is potentially helpful for the propagation of the contextual information learned at previous stages to the future ones to avoid redundant computation. Recurrent neural networks (RNNs) [69] are a type of neural networks constructed to extract contextual information from sequences, and thus have been extensively used in processing sequential data. With these concerns, we consider those $\mathcal{X}^{(k)}$s reconstructed by the totally K stages as a category of sequential data with length K, and then the recurrence of those stages can be naturally modeled by a RNN with recurrent hidden connections between PLs in (21).

Taken together, we consider three categories of contextual information in (21), i.e., the spatial information and spectral correlation in HS images, as well as the recurrence of the iterative stages, and employing a Transformer embedded convolutional recurrent neural network (CRNN) [70], [71] as the basic architecture of the prior-learning layers to uniformly exploit them. We now detailedly describe the structure of our prior-learning layers.

Our prior-learning layers consist of three categories of components:
(1) $Spatial$ unit ($1S$ unit): a vision Transformer unit. This unit only learn the spatial information.
(2) $Spatial + Stage$ unit ($2S$ unit): a 2D convolutional recurrent unit with recurrent hidden connections across iterative stages. This unit jointly learn the spatial information and the recurrence of the iterative stages.
(3) $Spatial + Stage + Spectral$ unit ($3S$ unit): a 3D convolutional recurrent unit with recurrent hidden connections over iterative stages. This unit jointly learn the spatial information, the spectral correlations and the recurrence of the iterative stages.

In every prior-learning layer $PL^{(k)}$ in the k-th stage, the residual learning network $\mathbf{Pnet}(\cdot)$ is composed of one $1S$ unit, one $2S$ unit and two $3S$ unit sequentially connected as follows:



(22)

then $\mathcal{X}^{(k)}$ can be easily calculated by $\mathcal{X}^{(k)} = \mathcal{Z}^{(k)} + \mathbf{Pnet}(\mathcal{Z}^{(k)})$ as in (20). Hereinafter, we use $\mathcal{H}_l^{(k)}$ to denote the outputed feature maps of the l-th unit of $\mathbf{Pnet}(\cdot)$ in the k-th stage. In the 3S unit, we adopt 3D convolution to extract the spatial and spectral features simultaneously, and meanwhile the recurrent hidden connections across stages are employed to achieve the information propagation learned in the last stage. The architecture of our 3S unit is shown in Fig. 1(b), where it is the l-th unit in the k-th stage with output $\mathcal{H}_l^{(k)}$. Considering that regular 3D convolution with $m \times m \times m$ filter can lead to a significant increase in parameters, we employ the separable 3D convolution (which splits the regular 3D filter $m \times m \times m$ into $1 \times m \times m$ and $m \times 1 \times 1$ to learn the spatial and spectral information respectively) as a alternative. It has been proven that the separable 3D convolution can achieve the same effect as the regular one with dramatically reduced memory usage and training time [44], [45], [34]. It can be seen that the 3S unit is designed as the weighted sum of two parts which apply separable 3D convolution respectively to two inputs, $\mathcal{H}_{l-1}^{(k)}$ for the spatial and spectral feature extraction and $\mathcal{H}_l^{(k-1)}$ for the information propagation across stages. We also adopt local residual connections in the second part to build a highway from $\mathcal{H}_l^{(k-1)}$ to $\mathcal{H}_l^{(k)}$ to facilitate information fusion and improve the optimization. We employ the weights

α and β to balance the importance of the two parts, and we set α and β learnable parameters of the network instead of fixed values (for example, $\alpha = 0.5$, $\beta = 0.5$). In the network training α and β are initialized to 0.9 and 0.1, respectively, to make the first part play a more important role in the beginning. In mathematical form, let $f_{s3dc}(\cdot)$ denote the separable 3D convolution, if the l-th unit in the k-th stage is a 3S unit, then its output $\mathcal{H}_l^{(k)}$ can been written as follows:

$$\begin{aligned}
&\mathcal{H}_{l\ \text{[feature extraction]}}^{(k)} = f_{s3dc}(\sigma(f_{s3dc}(\mathcal{H}_{l-1}^{(k)}))),\\
&\mathcal{H}_{l\ \text{[information propagation]}}^{(k)} = f_{s3dc}(\mathcal{H}_{l}^{(k-1)}) + \mathcal{H}_{l}^{(k-1)},\\
&\mathcal{H}_{l}^{(k)} = \alpha \cdot \mathcal{H}_{l\ \text{[feature extraction]}}^{(k)} + \beta \cdot \mathcal{H}_{l\ \text{[information propagation]}}^{(k)},
\end{aligned}$$

where $\sigma(x) = \max(x, 0)$ denotes the rectifier linear unit (ReLU) active function. In our $\mathbf{Pnet}(\cdot)$ in (22), given the current $\mathcal{Z}^{(k)} \in \mathbb{R}^{W\times H\times S}$, we first apply a 3S unit to extract shallow spatial and spectral feature $\mathcal{H}_1^{(k)}$, and then a 1S unit is sequentially employed to further extract deep spatial feature $\mathcal{H}_2^{(k)}$. In this way, the whole network can more focus on the spatial exploration under the condition that the spectral information can be extracted. It is worth mentioning that in the network training process with batch size B, the input of $\mathbf{Pnet}(\cdot)$ is of size $B\times W\times H\times S$, and it should be permuted and reshaped into five dimensions $B\times 1\times S\times W\times H$ before performing 3D convolution. In that case, the size of feature maps $\mathcal{H}_1^{(k)}$ is $B\times C\times S\times W\times H$, where C is the number of filters.

Swin Transformer [62] is a recently developed vision Transformer architecture and has shown great promise owing to its dual advantages in large-size image processing as CNN and long-range dependency modeling as Transformer. To effectively extract deep spatial feature from $\mathcal{H}_1^{(k)}$, we employ a Swin Transformer based structure in 1S unit. The architecture of our 1S unit is shown in Fig. 1(d), where it is the l-th unit in the k-th stage with output $\mathcal{H}_l^{(k)}$. We can see that 1S unit consists of three Consecutive Swin Transformer Blocks and a 2D convolutional layer, where the former extracts intermediate features and the latter further enhances its translational equivariance. As in Swin Transformer, we alternate the regular and shifted window partitioning strategy in Transformer layers to introduce cross-window connections while maintaining the efficient computation of non-overlapping windows. And for convenience, we denote two connected Swin Transformer layers with regular and shifted window partitioning respectively as a Consecutive Swin Transformer Block (CSTB), then a CSTB can be mathematically denoted as $f_{CSTB}(\cdot)$ and computed as

$$\begin{aligned}
\hat{\mathcal{H}}_1 &= \text{W-MSA}(\text{LN}(\mathcal{H})) + \mathcal{H},\\
\hat{\mathcal{H}}_2 &= \text{MLP}(\text{LN}(\hat{\mathcal{H}}_1)) + \hat{\mathcal{H}}_1,\\
\hat{\mathcal{H}}_3 &= \text{SW-MSA}(\text{LN}(\hat{\mathcal{H}}_2)) + \hat{\mathcal{H}}_2,\\
f_{CSTB}(\mathcal{H}) &= \text{MLP}(\text{LN}(\hat{\mathcal{H}}_3)) + \hat{\mathcal{H}}_3,
\end{aligned}$$

where $\mathcal{H}$ denotes the input; W-MSA and SW-MSA denote regular and shifted window partitioning based multi-head self-attention module, respectively; MLP and LN denote 2-layer MLP with GELU nonlinearity module and LayerNorm layer, respectively. We use f_{2dc} to denote 2D convolution, then our 1S unit can be written as

$$\mathcal{H}_l^{(k)} = f_{2dc}(f_{CSTB}(f_{CSTB}(f_{CSTB}(\mathcal{H}_{l-1}^{(k)})))).$$

It is remarkable that in (22), the input of 1S unit is $\mathcal{H}_1^{(k)}$ with size $B\times C\times S\times W\times H$, which should be reshaped in four dimensions to perform the Transformer and 2D convolution in 1S unit. In our framework we treat each band separately and integrate the channel B and S in $\mathcal{H}_1^{(k)}$ together, i.e., transform its size into $B*S\times C\times W\times H$ before feeding into 1S unit. Without causing any ambiguity, we still denote the input of 1S unit as $\mathcal{H}_1^{(k)}$.

At this point we extract shallow and deep features using 3S and 1S unit, respectively, and we then aggregate this two features to synthetically exploit the low and high frequency information. We achieve this point by sum operation and a 2S unit, see (22). Denote the features after summation as $\mathcal{H}_{\text{sum}}^{(k)}$ (in the case of (22), we have $\mathcal{H}_{\text{sum}}^{(k)} = \mathcal{H}_1^{(k)} + \mathcal{H}_2^{(k)}$), then the architecture of 2S unit can be seen in Fig. 1(c). We can see that 2S unit has a similar structure to the 3S unit except for the 2D convolution. The employment of 2S unit brings two benefits. On one hand, compared with 3D convolution, the 2D convolution can effectively integrate the shallow and deep spatial features in $\mathcal{H}_{\text{sum}}^{(k)}$ to pay more attention to spatial resolution, and significantly reduce the number of parameters. On the other hand, the recurrent connection from $\mathcal{H}_1^{(k-1)}$ helps the information propagation across stages to avoid redundant computation. In (22), the feature maps after 2S unit is $\mathcal{H}_3^{(k)}$ with size $B*S\times C\times W\times H$. $\mathcal{H}_3^{(k)}$ is then reshaped and permuted back into size $B\times C\times S\times W\times H$ and fed into a 3S unit to do the last fitting spatially and specially, and finally obtain the output $\mathbf{Pnet}(\mathcal{Z}^{(k)})$ with size $B\times W\times H\times S$. It is worth noting that in this 3S unit the second separated 3D convolutions in both of the two parts have only one filter to produce a proper $\mathcal{H}_4^{(k)}$ of size $B\times 1\times S\times W\times H$, and then $\mathcal{H}_4^{(k)}$ can be easily squeezed and permuted to the size $B\times W\times H\times S$, i.e., $\mathbf{Pnet}(\mathcal{Z}^{(k)})$.

Up to this point, we successfully construct an effective knowledge-driven, end-to-end and data-dependent HS image super-resolution framework based on deep unrolling technique and Transformer embedded convolutional recurrent neural network. We use Deep Unrolling based HS image Super-Resolution network (DUHSR) to denote the proposed framework. A concise illustration of its feed-forward structure is shown in Fig. 1. Next we will describe some details in the network training.

D. Network Training

Given the training data consisting of N low-high resolution HS image pairs $\{(\mathcal{Y}_1, \mathcal{X}_1), (\mathcal{Y}_2, \mathcal{X}_2), .., .(\mathcal{Y}_N, \mathcal{X}_N)\}$, the training loss of our network (21) with K stages is defined as follows:

$$\begin{aligned}
L(\Theta) = \frac{1}{N}\sum_{i=1}^{N}\Bigg(&\|\mathcal{X}_i^{(K)} - \mathcal{X}_i\|_F^2 + \lambda_1\|\mathcal{X}_i^{(K)} - \mathcal{X}_i\|_1\\
&+ \lambda_2\sum_{k=1}^{K-1}\|\mathcal{X}_i^{(k)} - \mathcal{X}_i\|_F^2 + \lambda_3\|\mathcal{E}_i^{(K)}\|_F^2\Bigg),
\end{aligned}$$

where $\|\cdot\|_1$ denotes the ℓ_1 norm, $\mathcal{X}^{(K)}$ and $\mathcal{X}^{(k)}$ are the final and per-stage outputs of the proposed network, λ_1, λ_2 and λ_3 are trade-off parameters, Θ is the learning parameters. In our experiments we fix $\lambda_1 = 1$, $\lambda_2 = 0.1$, $\lambda_3 = 0.01$ and $K = 5$. He initialization [72] is used to initialize the learning parameters and Adam [73] is used as the optimizer.

IV. Experiments

A. Compared Methods and Performance Evaluation Measures

We shall compare the performance of our propose procedure with several state-of-the-art image super-resolution approaches, including three nature image super-resolution ones SRCNN (super-resolution convolutional neural network) [30], EDSR (enhanced deep residual networks) [74], SwinIR [63], a model optimization based HS image super-resolution one NLRTV[1] [40], five deep learning based HS image super-resolution ones 3DFCNN (3d full convolutional neural networks) [33], MCNet (mixed convolutional network) [45], ERCSR [34], SFCSR[46], SSPSR [47], and a spatiospectral image super-resolution one US3RN[2] [56]. We also use Bicubic interpolation as the baseline method. To evaluate the quality of reconstructed HS images from the spatial and spectral perspectives, we employ totally seven assessments, including root mean square error (RMSE), peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), spectral angle mapper (SAM), relative dimensionless global error in synthesis (ERGAS), universal image quality index (UIQI) and degree of distortion (DD). The smaller values of RMSE, SAM, ERGAS and DD and the higher values of PSNR, SSIM and UIQI mean the better reconstruction performance.

B. Experimental Data Sets

We employ totally three different public data sets of hyperspectral images, including CAVE [75], Harvard [76] and Pavia Center [77]. The CAVE data set contains totally 32 hyperspectral images of a wide variety of real-world materials and objects, where each image is of spatial size 512×512 and acquired using 31 spectral bands ranging from 400nm to 700nm. In order to generate the low-high resolution image pairs, we take the original images with size $512 \times 512 \times 31$ as the high resolution images (ground truth), and then spatially downsample them with $factor = 2$ and $factor = 4$ respectively to generate the corresponding low resolution images. The downsampling operation is implemented by averaging over the pixel blocks of size $factor \times factor$ as done in [78], [79]. Then, we randomly select 20 images from the totally 32 ones and extract overlapped image patches of size 96×96 from them for training, and thus the size of high resolution images of training samples is $96 \times 96 \times 31$ and the corresponding low resolution images are of size $48 \times 48 \times 31$ $(factor = 2)$ or $24 \times 24 \times 31$ $(factor = 4)$. We use the remaining 12 images as the test samples with high resolution images of $512 \times 512 \times 31$ and the corresponding low resolution ones $256 \times 256 \times 31$ $(factor = 2)$ or $128 \times 128 \times 31$ $(factor = 4)$.

TABLE I
Quantitative results of the tested methods on the CAVE data set.

Methods	$factor = 2$						
	RMSE	PSNR	SSIM	SAM	ERGAS	UIQI	DD
Bicubic	4.158	36.516	0.969	4.440	9.693	0.837	1.510
SRCNN	2.695	40.288	0.981	4.510	6.696	0.859	1.050
EDSR	1.978	43.085	0.987	4.331	4.892	0.865	0.796
SwinIR	1.861	43.665	0.989	3.683	4.633	0.885	0.683
3DFCNN	2.395	41.306	0.984	4.466	5.814	0.858	0.935
MCNet	1.865	43.641	0.988	3.709	4.591	0.883	0.698
ERCSR	1.798	44.001	0.989	3.615	4.457	0.889	0.666
SSPSR	1.783	43.901	0.989	3.715	4.427	0.890	0.693
SFCSR	1.798	44.030	0.989	3.608	4.451	0.887	0.666
US3RN	2.275	41.775	0.984	4.325	5.630	0.866	0.891
OURS	**1.666**	**44.532**	**0.990**	**3.563**	**4.261**	**0.894**	**0.624**
Methods	$factor = 4$						
	RMSE	PSNR	SSIM	SAM	ERGAS	UIQI	DD
Bicubic	8.239	30.555	0.904	5.938	9.094	0.707	3.274
SRCNN	5.023	34.942	0.940	5.885	5.808	0.776	2.024
EDSR	4.810	35.187	0.884	11.433	5.664	0.666	2.562
SwinIR	4.058	37.034	0.961	4.888	4.744	0.797	1.471
3DFCNN	4.619	35.769	0.950	5.452	5.320	0.783	1.826
MCNet	4.044	37.039	0.960	4.694	4.661	0.802	1.482
ERCSR	3.936	37.316	0.963	4.622	4.612	0.807	1.402
SSPSR	3.786	37.579	0.964	4.650	**4.381**	0.810	1.421
SFCSR	3.907	37.424	0.963	4.604	4.583	0.807	1.396
US3RN	4.467	35.972	0.954	5.257	5.144	0.787	1.721
OURS	**3.655**	**37.902**	**0.966**	**4.575**	**4.381**	**0.811**	**1.312**

The Harvard data set consists of 50 hyperspectral images of real-world indoor and outdoor scenes under daylight illumination, each of which is of spatial size 1040×1392 and captured using 31 spectral bands ranging from 420nm to 720nm. We randomly select 20 images from the Harvard data set for training, and the rest for testing. We process them in a similar way as done for the CAVE data set, giving the training data with high resolution images of size $96 \times 96 \times 31$ and the corresponding low resolution images of size $48 \times 48 \times 31$ $(factor = 2)$ or $24 \times 24 \times 31$ $(factor = 4)$, and the test data with high resolution images of size $1040 \times 1392 \times 31$ and the corresponding low resolution ones $520 \times 696 \times 31$ $(factor = 2)$ or $260 \times 348 \times 31$ $(factor = 4)$.

The Pavia Center data set is a real-world data acquired by the reflective optics system imaging spectrometer optical sensor during a flight campaign over Pavia, Italy. This image data is of size $1096 \times 715 \times 93$ with a spatial resolution of 1.3m. We use the subgraph of size $256 \times 256 \times 93$ in its upper left corner as the test image, and the rest for training. We process such training and test image data in a similar way to the previous two data sets, which gives the training data with high resolution images of size $96 \times 96 \times 93$ and the corresponding low resolution images of size $48 \times 48 \times 93$ $(factor = 2)$ or $24 \times 24 \times 93$ $(factor = 4)$, and the test data with high resolution images of size $256 \times 256 \times 93$ and the corresponding low resolution ones $128 \times 128 \times 93$ $(factor = 2)$ or $64 \times 64 \times 93$ $(factor = 4)$.

[1]Due to its high computational complexity and time cost, NLRTV is difficult to be applied to CAVE and Harvard data sets with many large-size test images, and thus we only compare with it in the Pavia Center data set.

[2]US3RN is originally designed for the joint spatial and spectral super-resolution problem, and its model combines the two auxiliary tasks with a trade-off parameter α. In our experiments, we set $\alpha = 1$ for US3RN, and in that case it degenerates to only introduce the single HS image super-resolution auxiliary task.

C. Comparisons With State-of-the-Arts

In this subsection, we compare the reconstruction results of the aforementioned methods. Table I lists the average RMSE, PSNR, SSIM, SAM, ERGAS, UIQI and DD values over the 12 testing images of the CAVE data set recovered by those compared methods, where the best results are marked in bold. As we can see from Table I, our method performs better among all the compared methods in terms of the seven assessments, which shows its superiority over the other ones. More precisely, the simple interpolation method, Bicubic, yields quite poor result, while SRCNN achieves significantly better performance through the employment of convolutional neural network, and furthermore, by introducing 3d convolutional structures, 3DFCNN has further improved the reconstruction result of SRCNN from both spatial and spectral perspectives. EDSR achieves better spatial reconstruction result through deeper network structures and larger number of parameters, and further, SwinIR greatly surpasses EDSR owing to the advantage of vision Transformer to model long-range dependency with the shifted window scheme. However, due to only considering spatial information and ignoring the underlying spectral correlations, the spectral fidelity of EDSR and SwinIR is not satisfactory. By simultaneously exploiting the spatial and spectral information of the HS images, the four recently developed single HS image super-resolution methods, MCNet, ERCSR, SSPSR and SFCSR, achieve good reconstruction accuracy in spatial and spectral perspective. Meanwhile, we can see that the performance of US3RN is not very well, this is partially because that this method is specially designed for the combined task and thus cannot well handle the auxiliary simplex one.

Clearly, among all the compared methods, our method cannot only achieve the best spatial reconstruction accuracy, but also get significantly superior results to the competing methods in terms of SAM and DD that describe the spectral similarity between the recovered HS images and the ground-truth ones, which proves the superiority of our method in spectral information reconstruction and spectral distortion prevention. This owes a great deal to the knowledge driven deep unrolling framework learning the implicit priors and the convolutional recurrent neural network modeling the spectral corrections of the HS images. Furthermore, to compare the performance of those methods on each testing image, we demonstrate the PSNR, ERGAS, SSIM, UIQI, SAM and DD curves as functions of the indices of the 12 testing images in the supplementary material.

To further intuitively compare the performance of those compared methods, Fig. 2 shows the reconstructed images and corresponding error images of the test image 'chart and stuffed toy' for the case of $factor = 2$ and test image 'clay' for the case of $factor = 4$ in CAVE data set. It can be seen from this figure that our method obviously outperforms other competing methods in recovering the image details, which proves its effectiveness in the reconstruction of spatial information. Meanwhile, we show the spectral curves at two pixels in the reconstructed image of the test image 'chart and stuffed toy' for the case of $factor = 2$ and test image 'clay' for the case of $factor = 4$ in CAVE data set in Fig. 3, from which one can see our method effectively prevents spectral distortion, and thus better reconstruct the spectral information of the images.

TABLE II
QUANTITATIVE RESULTS OF THE TESTED METHODS ON THE HARVARD DATA SET.

Methods	$factor = 2$						
	RMSE	PSNR	SSIM	SAM	ERGAS	UIQI	DD
Bicubic	2.948	39.402	0.957	2.899	6.151	0.793	1.543
SRCNN	2.151	42.094	0.969	2.853	5.630	0.811	1.205
EDSR	1.727	43.928	0.973	3.125	5.792	0.831	1.504
SwinIR	1.724	44.176	0.977	2.803	4.521	0.836	1.009
3DFCNN	1.926	43.092	0.975	2.753	4.609	0.831	1.081
MCNet	1.671	44.298	0.978	**2.646**	4.131	0.844	0.958
ERCSR	1.681	44.428	0.978	2.693	4.257	0.844	0.966
SSPSR	1.669	44.448	0.978	2.723	4.297	0.841	0.965
SFCSR	1.662	44.498	0.978	2.695	4.236	0.845	0.958
US3RN	1.760	43.965	0.977	2.786	4.670	0.830	1.015
OURS	**1.599**	**44.817**	**0.979**	2.681	**4.115**	**0.846**	**0.927**
Methods	$factor = 4$						
	RMSE	PSNR	SSIM	SAM	ERGAS	UIQI	DD
Bicubic	5.641	33.773	0.890	3.441	5.307	0.626	2.819
SRCNN	3.834	37.212	0.921	3.260	4.172	0.707	2.044
EDSR	3.467	38.178	0.931	3.236	3.540	0.732	1.864
SwinIR	3.443	38.425	0.931	3.218	3.522	0.735	1.845
3DFCNN	3.711	37.535	0.928	3.265	3.733	0.721	1.951
MCNet	3.355	38.504	0.935	3.125	3.345	0.741	1.766
ERCSR	3.363	38.623	0.934	3.193	3.457	0.739	1.789
SSPSR	3.386	38.580	0.934	3.307	3.534	0.738	1.836
SFCSR	3.383	38.572	0.933	3.254	3.570	0.738	1.812
US3RN	3.455	38.349	0.931	3.351	3.661	0.728	1.857
OURS	**3.237**	**38.997**	**0.937**	**3.110**	**3.336**	**0.747**	**1.722**

Table II shows the quantitative results of the Harvard data set, from which one can see that the proposed method is superior to the competitors numerically. As done on the CAVE data set, we plot the assessment curves as functions of the indices of the 30 testing images in the supplementary material. We further show in Fig. 4 that the reconstructed images and corresponding error images of the test image 'imgc9' for the case of $factor = 2$ and test image 'imgc4' for the case of $factor = 4$, and meanwhile the spectral curves at two pixels in the reconstructed images of the above two test images are displayed in Fig. 5. It is easy to see from Figs. **??**-5 that our method outperforms the competing one in recovering the spatial and spectral information.

The quantitative results of all the compared methods on the Pavia Center data set are also shown in Table III. In addition, we display the corresponding reconstructed and error images for the $91th$ and $80th$ bands of those methods in Fig. 6 for the intuitive comparison. And the spectral curves at two pixels in the reconstructed image are also shown in Fig. 7. It can be observed from Table III and Figs. 6-7 that the proposed knowledge-driven deep network get the best reconstruction results, both in terms of quantitative assessments and recovering image details.

D. Model Complexity and Ablation Study

We further compare the model complexity of the aforementioned methods from the aspect of computational time and the number of parameters, and explore the impact of unrolling

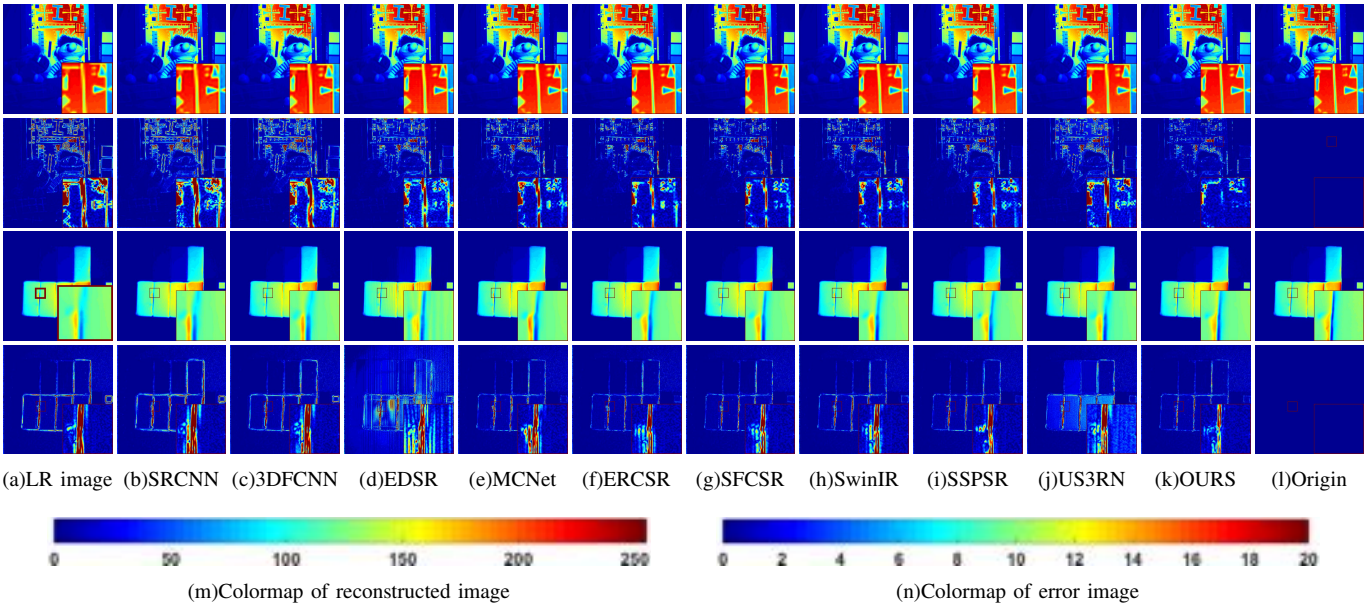


Fig. 2. The first two lines are the reconstructed images and corresponding error images of the test image 'chart and stuffed toy' in CAVE for the case $factor = 2$, and the last two lines are the reconstructed images and corresponding error images of the test image 'clay' in CAVE for the case $factor = 4$, respectively.(a) LR-HS image; (b) SRCNN; (c) 3DFCNN; (d) EDSR; (e) MCNet; (f) ERCSR; (g) SFCSR; (h) SwinIR; (i) SSPSR; (j) US3RN; (k)Our Method; (l) Ground truth; (m) Color map of the reconstructed images; (n) Color map of the error images.

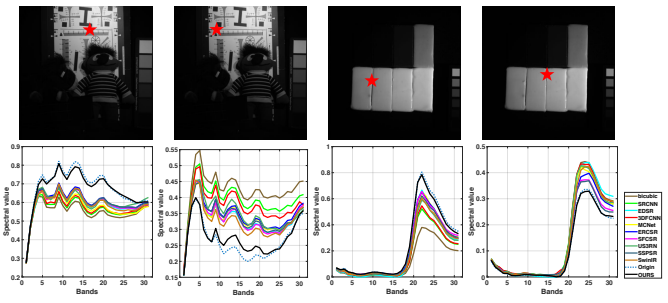


Fig. 3. The spectral curves at two pixels in the reconstructed image of the test image 'chart and stuffed toy' for the case $factor = 2$ and test image 'clay' for the case $factor = 4$ in CAVE data set. The top line is the location of the pixel in the image, and the bottom line is the corresponding spectral curves.

TABLE III
QUANTITATIVE RESULTS OF THE TESTED METHODS ON THE PAVIA CENTER DATA SET.

Methods	$factor = 2$						
	RMSE	PSNR	SSIM	SAM	ERGAS	UIQI	DD
Bicubic	4.704	34.681	0.928	1.687	3.912	0.956	2.993
SRCNN	4.285	35.491	0.942	1.604	3.567	0.960	2.720
EDSR	3.728	36.702	0.955	1.465	3.101	0.971	2.328
SwinIR	3.418	37.455	0.962	1.399	2.846	0.975	2.004
3DFCNN	4.158	35.752	0.945	1.604	3.460	0.963	2.610
MCNet	3.616	36.965	0.958	1.435	3.007	0.972	2.162
ERCSR	3.539	37.154	0.960	1.449	2.945	0.973	2.095
SSPSR	3.395	37.514	0.962	1.391	2.832	0.975	2.049
SFCSR	3.615	36.968	0.959	1.464	3.010	0.972	2.125
US3RN	3.169	38.113	0.966	**1.351**	2.646	0.978	1.924
NLRTV	3.870	36.377	0.950	1.920	3.185	0.970	2.645
OURS	**3.052**	**38.441**	**0.968**	1.428	**2.536**	**0.981**	**1.878**
Methods	$factor = 4$						
	RMSE	PSNR	SSIM	SAM	ERGAS	UIQI	DD
Bicubic	8.608	29.433	0.752	2.611	3.544	0.807	5.855
SRCNN	8.278	29.772	0.776	2.560	3.413	0.837	5.599
EDSR	7.714	30.385	0.806	2.475	3.185	0.861	5.071
SwinIR	7.461	30.674	0.823	2.408	3.082	0.871	4.869
3DFCNN	8.092	29.970	0.785	2.516	3.339	0.845	5.415
MCNet	7.658	30.448	0.810	2.419	3.166	0.862	4.865
ERCSR	7.588	30.529	0.817	2.527	3.134	0.867	4.776
SSPSR	7.223	30.956	0.834	2.433	2.996	0.878	4.482
SFCSR	7.427	30.715	0.821	2.499	3.077	0.869	4.700
US3RN	8.515	29.527	0.771	2.871	3.519	0.832	5.665
NLRTV	7.798	30.291	0.808	3.615	3.185	0.869	5.276
OURS	**6.740**	**31.557**	**0.856**	**2.397**	**2.793**	**0.899**	**4.206**

stages K on our model performance and the effectiveness of the various modules in our framework, including 3D convolution, vision Transformer and the recurrent hidden connections over iterative stages. We show the results in the supplementary material due to page limitations.

V. CONCLUSION

In this paper, we present an effective hyperspectral image super-resolution framework based on knowledge-driven deep unrolling and Transformer embedded convolutional recurrent neural network. We first build a maximum a posterior based energy model with implicit priors, and then solve it by alternating optimization to determine an elementary iteration mechanism. we unroll such iteration mechanism with an ingenious convolutional recurrent neural network where the implicit spatial-spectral priors and recurrence of the iterative reconstruction stages can be modeled and learned by 3D convolution, vision Transformer and the recurrent hidden connections across iterations. As a consequence, we successfully construct an effective knowledge-driven, end-to-end and data-dependent HS image super-resolution framework. Extensive experiments are conducted on three different data sets to demonstrate that the proposed framework can achieve smaller

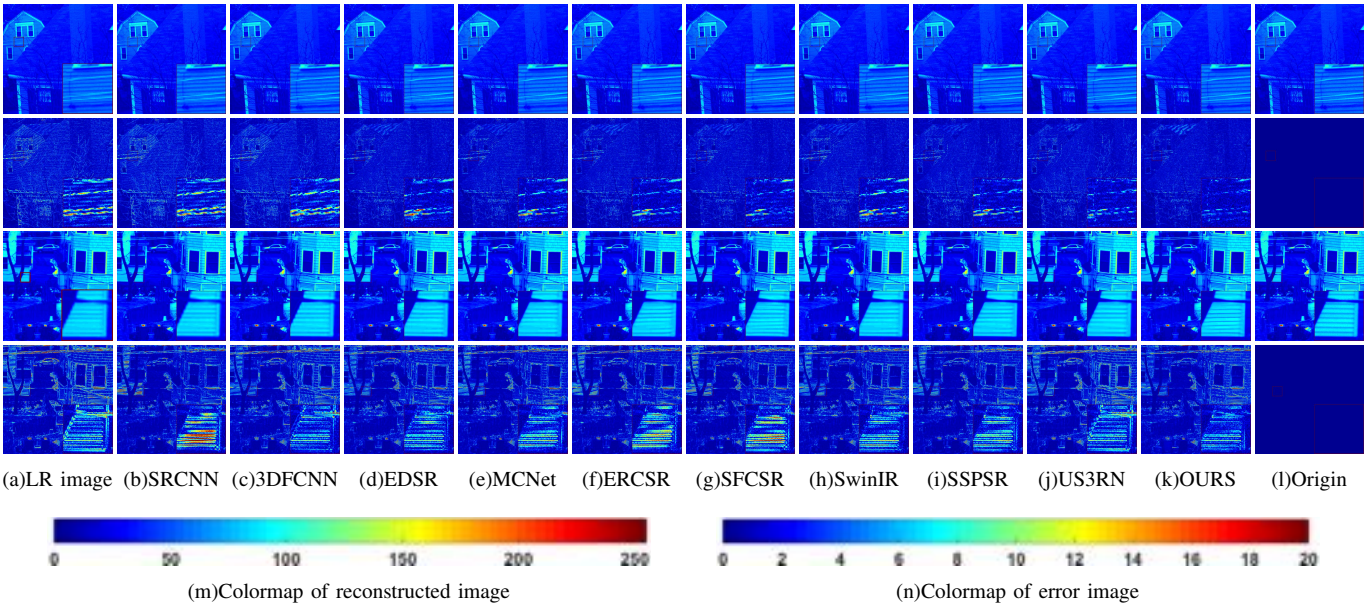


Fig. 4. The first two lines are the reconstructed images and corresponding error images of the test image 'imgc9' in Harvard for the case $factor = 2$, and the last two lines are the reconstructed images and corresponding error images of the test image 'imgc4' in Harvard for the case $factor = 4$, respectively.(a) LR-HS image; (b) SRCNN; (c) 3DFCNN; (d) EDSR; (e) MCNet; (f) ERCSR; (g) SFCSR; (h) SwinIR; (i) SSPSR; (j) US3RN; (k)Our Method; (l) Ground truth; (m) Color map of the reconstructed images; (n) Color map of the error images.

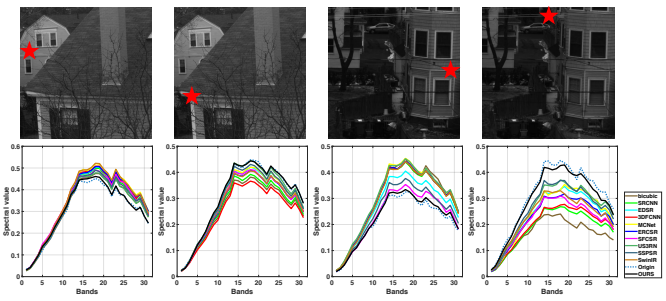


Fig. 5. The spectral curves at two pixels in the reconstructed image of the test image 'imgc9' for the case $factor = 2$ and test image 'imgc4' for the case $factor = 4$ in Harvard data set. The top line is the location of the pixel in the image, and the bottom line is the corresponding spectral curves.

reconstruction errors and better visual quality than several state-of-the-art HS image super-resolution methods.

It should be finally pointed out that two interesting directions need to be further investigated in the future. On one hand, as most of the previous methods, the proposed model uses linear mappings to model the relations between the low and high resolution HS image pairs, while in real cases such relations are always non-linear due to intensity alignment and other operations. And it could be valuable to incorporate the nonlinearity into our model to further improve its performance. On the other hand, the proposed model can be treated as a general framework and thus expanded to other related tasks, such as video super-resolution, snapshot compressive imaging, and so on.

REFERENCES

[1] X. Wu, D. Hong, J. Tian, J. Chanussot, W. Li, and R. Tao, "ORSIm Detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 5146–5158, 2019.

[2] J. Yang, Y.-Q. Zhao, J. C.-W. Chan, and S. G. Kong, "Coupled sparse denoising and unmixing with low-rank constraint for hyperspectral image," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 3, pp. 1818–1833, 2016.

[3] H. Van Nguyen, A. Banerjee, and R. Chellappa, "Tracking via object reflectance using a hyperspectral video camera," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 44–51.

[4] X. Cao, F. Zhou, L. Xu, D. Meng, Z. Xu, and J. Paisley, "Hyperspectral image segmentation with markov random fields and a convolutional neural network," *IEEE Transactions on Image Processing*, vol. PP, no. 99, p. 2354, 2017.

[5] D. Lorente, N. Aleixos, J. Gómez-Sanchis, S. Cubero, O. L. García-Navarrete, and J. Blasco, "Recent advances and applications of hyperspectral imaging for fruit and vegetable quality assessment," *Food and Bioprocess Technology*, vol. 5, no. 4, pp. 1121–1142, 2012.

[6] G. A. Shaw and H. K. Burke, "Spectral imaging for remote sensing," *Lincoln Laboratory Journal*, vol. 14, no. 1, pp. 3–28, 2003.

[7] Y. Gu, Y. Zhang, and J. Zhang, "Integration of spatial–spectral information for resolution enhancement in hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 5, pp. 1347–1358, 2008.

[8] Gemine, Vivone, Rocco, Restaino, Jocelyn, and Chanussot, "A regression-based high-pass modulation pansharpening approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 984–996, 2018.

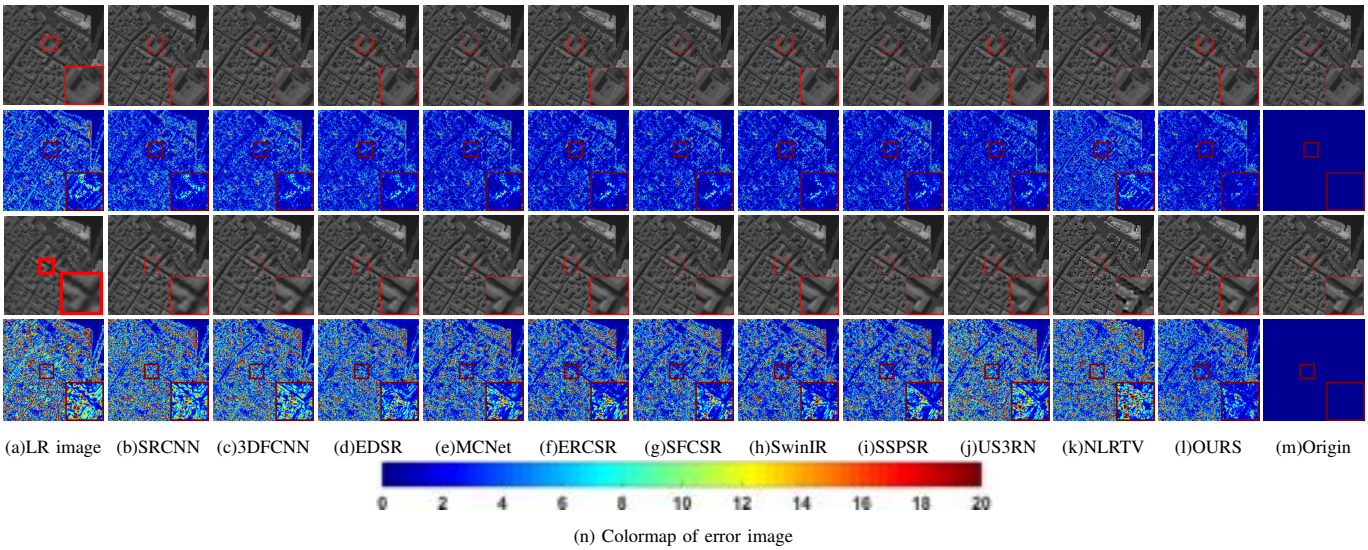
[9] L. Loncan, L. B. De Almeida, J. M. Bioucas-Dias, X. Briottet, J. Chanussot, N. Dobigeon, S. Fabre, W. Liao, G. A. Licciardi, M. Simoes *et al.*, "Hyperspectral pansharpening: A review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 3, no. 3, pp. 27–46, 2015.

[10] X. X. Zhu and R. Bamler, "A sparse image fusion algorithm with application to pan-sharpening," *IEEE Transactions on Geoscience and Remote sensing*, vol. 51, no. 5, pp. 2827–2836, 2013.

[11] B. Aiazzi, S. Baronti, and M. Selva, "Improving component substitution pansharpening through multivariate regression of ms + pan data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 10, pp. 3230–3239, 2007.

[12] A. Garzelli and F. Nencini, "Interband structure modeling for pan-sharpening of very high-resolution multispectral images," *Information Fusion*, vol. 6, no. 3, pp. 213–224, 2005.

[13] W. Huang, L. Xiao, Z. Wei, H. Liu, and S. Tang, "A new pan-sharpening



(n) Colormap of error image

Fig. 6. The first two lines are the reconstructed images and corresponding error images of the $91th$ band in the test image of Pavia Center for the case $factor = 2$, and the last two lines are the reconstructed images and corresponding error images of the $80th$ band for the case $factor = 4$, respectively.(a) LR-HS image; (b) SRCNN; (c) 3DFCNN; (d) EDSR; (e) MCNet; (f) ERCSR; (g) SFCSR; (h) SwinIR; (i) SSPSR; (j) US3RN; (k) NLRTV; (l)Our Method; (m) Ground truth; (n) Color map of the error images.

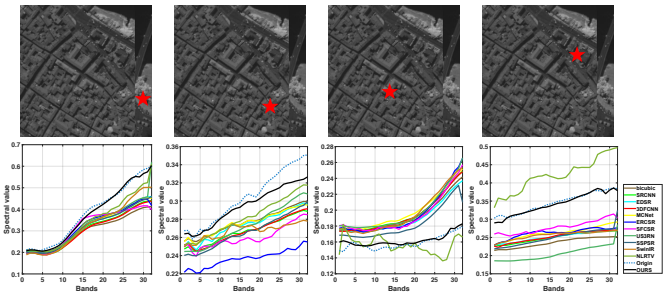


Fig. 7. The spectral curves at two pixels in the reconstructed image of the test image for the case $factor = 2$ and $factor = 4$ in Pavia Center data set. The top line is the location of the pixel in the image, and the bottom line is the corresponding spectral curves.

method with deep neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 5, pp. 1037–1041, 2015.

[14] P. Guan and E. Y. Lam, "Multistage dual-attention guided fusion network for hyperspectral pansharpening," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[15] R. Dian, L. Fang, and S. Li, "Hyperspectral image super-resolution via non-local sparse tensor factorization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5344–5353.

[16] S. Li, R. Dian, L. Fang, and J. M. Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4118–4130, 2018.

[17] R. Dian, S. Li, and L. Fang, "Learning a low tensor-train rank representation for hyperspectral image super-resolution," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2672–2683, 2019.

[18] R. Dian, S. Li, A. Guo, and L. Fang, "Deep hyperspectral image sharpening," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 11, pp. 5345–5355, 2018.

[19] R. Dian and S. Li, "Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 5135–5146, 2019.

[20] R. Dian, S. Li, L. Fang, T. Lu, and J. M. Bioucas-Dias, "Nonlocal sparse tensor factorization for semiblind hyperspectral and multispectral image fusion," *IEEE transactions on cybernetics*, vol. 50, no. 10, pp. 4469–4480, 2019.

[21] R. Dian, S. Li, and X. Kang, "Regularizing hyperspectral and multispectral image fusion by cnn denoiser," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 3, pp. 1124–1135, 2020.

[22] K. Wang, Y. Wang, X.-L. Zhao, J. C.-W. Chan, Z. Xu, and D. Meng, "Hyperspectral and multispectral image fusion via nonlocal low-rank tensor decomposition and spectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 7654–7671, 2020.

[23] Y. Yuan, X. Zheng, and X. Lu, "Hyperspectral image superresolution by transfer learning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 5, pp. 1963–1974, 2017.

[24] Z. Guo, T. Wittman, and S. Osher, "L1 unmixing and its application to hyperspectral image enhancement," in *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XV*, 2009.

[25] W. Huang, L. Xiao, H. Liu, and Z. Wei, "Hyperspectral imagery super-resolution by compressive sensing inspired dictionary learning and spatial-spectral regularization," *Sensors*, vol. 15, no. 1, pp. 2041–2058, 2015.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residuallearning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[28] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *arXiv preprint arXiv:1902.06068*, 2019.

[29] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, "Deep learning for single image super-resolution: A brief review," *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3106–3121, 2019.

[30] C. Dong, C. L. Chen, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–302, 2016.

[31] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, and Z. Wang, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4681–4690.

[32] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3147–3155.

[33] S. Mei, X. Yuan, J. Ji, Y. Zhang, S. Wan, and Q. Du, "Hyperspectral image spatial super-resolution via 3d full convolutional neural network," *Remote Sensing*, vol. 9, no. 11, p. 1139, 2017.

[34] Q. Li, Q. Wang, and X. Li, "Exploring the relationship between 2d/3d convolution for hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 10, pp. 8693–8703, 2021.

[35] D. Liu, J. Li, and Q. Yuan, "A spectral grouping and attention-driven residual dense network for hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 9, pp. 7711–7725, 2021.

[36] T. Akgun, Y. Altunbasak, and R. M. Mersereau, "Super-resolution reconstruction of hyperspectral images," *IEEE Transactions on Image Processing*, vol. 14, no. 11, pp. 1860–1875, 2005.

[37] H. Huang, A. G. Christodoulou, and W. Sun, "Super-resolution hyperspectral imaging with unknown blurring by low-rank and group-sparse modeling," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 2155–2159.

[38] X. Xu, X. Tong, J. Li, H. Xie, Y. Zhong, L. Zhang, and D. Song, "Hyperspectral image super resolution reconstruction with a joint spectral-spatial sub-pixel mapping model," in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2016, pp. 6129–6132.

[39] S. He, H. Zhou, Y. Wang, W. Cao, and Z. Han, "Super-resolution reconstruction of hyperspectral images via low rank tensor modeling and total variation regularization," *International Geoscience and Remote Sensing Symposium*, pp. 6962–6965, 2016.

[40] Y. Wang, X. Chen, Z. Han, and S. He, "Hyperspectral image super-resolution via nonlocal low-rank tensor approximation and total variation regularization," *Remote Sensing*, vol. 9, no. 12, p. 1286, 2017.

[41] H. Irmak, G. B. Akar, and S. E. Yuksel, "A map-based approach for hyperspectral imagery super-resolution," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2942–2951, 2018.

[42] J. Li, Q. Yuan, H. Shen, X. Meng, and L. Zhang, "Hyperspectral image super-resolution by spectral mixture analysis and spatial–spectral group sparsity," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 9, pp. 1250–1254, 2016.

[43] Y. Yuan, X. Zheng, and X. Lu, "Hyperspectral image superresolution by transfer learning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 5, pp. 1963–1974, 2017.

[44] Q. Wang, Q. Li, and X. Li, "Spatial-spectral residual network for hyperspectral image super-resolution," *arXiv preprint arXiv:2001.04609*, 2020.

[45] Q. Li, Q. Wang, and X. Li, "Mixed 2d/3d convolutional network for hyperspectral image super-resolution," *Remote sensing*, vol. 12, no. 10, p. 1660, 2020.

[46] Q. Wang, Q. Li, and X. Li, "Hyperspectral image superresolution using spectrum and feature context," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 11, pp. 11 276–11 285, 2020.

[47] J. Jiang, H. Sun, X. Liu, and J. Ma, "Learning spatial-spectral prior for super-resolution of hyperspectral imagery," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1082–1096, 2020.

[48] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *IEEE Signal Processing Magazine*, vol. 38, no. 2, pp. 18–44, 2021.

[49] S. Markowitz, C. Snyder, Y. C. Eldar, and M. N. Do, "Multimodal unrolled robust pca for background foreground separation," *IEEE Transactions on Image Processing*, 2022.

[50] H. Van Luong, B. Joukovsky, and N. Deligiannis, "Designing interpretable recurrent neural networks for video reconstruction via deep unfolding," *IEEE Transactions on Image Processing*, vol. 30, pp. 4099–4113, 2021.

[51] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proceedings of the 27th international conference on international conference on machine learning*, 2010, pp. 399–406.

[52] R. Liu, L. Ma, J. Zhang, X. Fan, and Z. Luo, "Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 561–10 570.

[53] Q. Xie, M. Zhou, Q. Zhao, Z. Xu, and D. Meng, "Mhf-net: An interpretable deep network for multispectral and hyperspectral image fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[54] W. Dong, C. Zhou, F. Wu, J. Wu, G. Shi, and X. Li, "Model-guided deep hyperspectral image super-resolution," *IEEE Transactions on Image Processing*, vol. 30, pp. 5754–5768, 2021.

[55] T. Huang, W. Dong, X. Yuan, J. Wu, and G. Shi, "Deep gaussian scale mixture prior for spectral compressive imaging," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 216–16 225.

[56] Q. Ma, J. Jiang, X. Liu, and J. Ma, "Deep unfolding network for spatiospectral image super-resolution," *IEEE Transactions on Computational Imaging*, vol. 8, pp. 28–40, 2021.

[57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[58] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[59] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," *arXiv preprint arXiv:2105.05537*, 2021.

[60] G. Sun, Y. Liu, T. Probst, D. P. Paudel, N. Popovic, and L. Van Gool, "Boosting crowd counting with transformers," *arXiv preprint arXiv:2105.10926*, 2021.

[61] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 299–12 310.

[62] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[63] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1833–1844.

[64] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE transactions on Image Processing*, vol. 4, no. 7, pp. 932–946, 1995.

[65] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep cnn denoiser prior for image restoration," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3929–3938.

[66] R. Liu, Z. Jiang, X. Fan, and Z. Luo, "Knowledge-driven deep unrolling for robust image layer separation," *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

[67] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[68] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4509–4522, 2017.

[69] J. L. Elman, "Finding structure in time," *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.

[70] Z. Zuo, B. Shuai, G. Wang, X. Liu, X. Wang, B. Wang, and Y. Chen, "Convolutional recurrent neural networks: Learning spatial dependencies for image representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 18–26.

[71] C. Qin, J. Schlemper, J. Caballero, A. N. Price, J. V. Hajnal, and D. Rueckert, "Convolutional recurrent neural networks for dynamic mr image reconstruction," *IEEE Transactions on Medical Imaging*, vol. 38, no. 1, pp. 280–290, 2018.

[72] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.

[73] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[74] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.

[75] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, "Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum," *IEEE transactions on image processing*, vol. 19, no. 9, pp. 2241–2253, 2010.

[76] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *CVPR 2011*. IEEE, 2011, pp. 193–200.

[77] F. Dell'Acqua, P. Gamba, A. Ferrari, J. A. Palmason, J. A. Benediktsson, and K. Árnason, "Exploiting spectral and spatial information in hyperspectral urban data with high resolution," *IEEE Geoscience and Remote Sensing Letters*, vol. 1, no. 4, pp. 322–326, 2004.

[78] R. Kawakami, Y. Matsushita, J. Wright, M. Ben-Ezra, Y.-W. Tai, and K. Ikeuchi, "High-resolution hyperspectral imaging via matrix factorization," in *CVPR 2011*. IEEE, 2011, pp. 2329–2336.

[79] N. Akhtar, F. Shafait, and A. Mian, "Sparse spatio-spectral representation for hyperspectral image super-resolution," in *European conference on computer vision*. Springer, 2014, pp. 63–78.