Pixel Adaptive Deep Unfolding Transformer for Hyperspectral Image Reconstruction

Miaoyu Li[1], Ying Fu[1]*, Ji Liu[2], Yulun Zhang[3]
[1]Beijing Institute of Technology, [2]Baidu Inc., [3]ETH Zürich
miaoyli@bit.edu.cn, fuying@bit.edu.cn, liuji04@baidu.com, yulun100@gmail.com

Abstract

Hyperspectral Image (HSI) reconstruction has made gratifying progress with the deep unfolding framework by formulating the problem into a data module and a prior module. Nevertheless, existing methods still face the problem of insufficient matching with HSI data. The issues lie in three aspects: 1) fixed gradient descent step in the data module while the degradation of HSI is agnostic in the pixel-level. 2) inadequate prior module for 3D HSI cube. 3) stage interaction ignoring the differences in features at different stages. To address these issues, in this work, we propose a Pixel Adaptive Deep Unfolding Transformer (PADUT) for HSI reconstruction. In the data module, a pixel adaptive descent step is employed to focus on pixel-level agnostic degradation. In the prior module, we introduce the Non-local Spectral Transformer (NST) to emphasize the 3D characteristics of HSI for recovering. Moreover, inspired by the diverse expression of features in different stages and depths, the stage interaction is improved by the Fast Fourier Transform (FFT). Experimental results on both simulated and real scenes exhibit the superior performance of our method compared to state-of-the-art HSI reconstruction methods. The code is released at: https://github.com/MyuLi/PADUT

1. Introduction

Hyperspectral Images (HSIs) have been widely applied in multiple fields, *e.g.*, material identification [16, 17], spectral unmixing [15], and medical analysis [46, 7]. Conventional hyperspectral imagers are time-consuming and inflexible when scanning a scene to capture the target HSI. To overcome this limitation and capture the diverse reluctance of a real scene, coded aperture snapshot spectral imaging (CASSI) modulates 3D HSI into a compressed 2D measurement by a coded aperture and a disperser [30]. Since the image quality of desired 3D HSI is restricted by the performance of reconstruction algorithms, exploring effective reconstruction algorithms is of much importance.

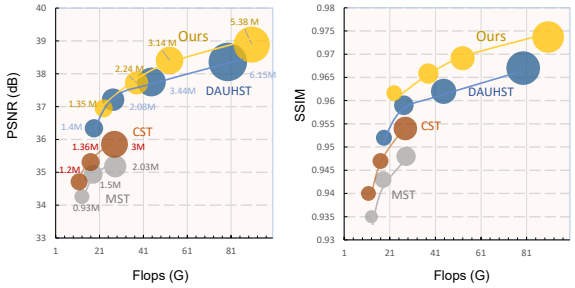

Figure 1: Performance vs. GFLOPs. Under different model sizes, PADUT outperforms the state-of-the-art methods.

To address the ill-posed reverse problem of HSI reconstruction, traditional model-based methods explore various priors in different solution space with interpretability. The widely employed priors can be summarized as non-local similarity [9, 43], low-rank property [22], sparsity [41] and total variation [6, 37], *etc*. However, these hand-crafted priors have limited generalization ability and often result in a mismatch between prior assumptions and the problem. Consequently, such methods cannot characterize the features of HSI under various scenarios and typically require time-consuming numerical iteration.

Recently, deep convolutional neural networks [36, 26, 3, 24] have been applied for HSI reconstruction and achieved decent performance. According to different learning strategies, deep learning-based methods can be broadly categorized into two groups, *i.e.*, end-to-end learning methods [36, 24] and model-aided methods [34, 4]. End-to-end learning methods recover the original HSI via brute-force mapping to learn the spatial and spectral information. Some of these typical works include λ-Net [26], TSA-Net [24], and MST [3]. Without the guidance of physical models, these end-to-end learning methods are black boxes and lack transparency. In contrast, the model-aided methods [34] leverage the physical characteristics of HSI degradation into

*Corresponding Author

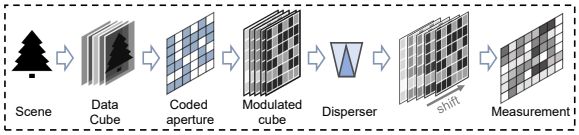

Figure 2: The coded aperture snapshot spectral imaging (CASSI) system. Different positions in the coded measurement may suffer from different levels of degradation.

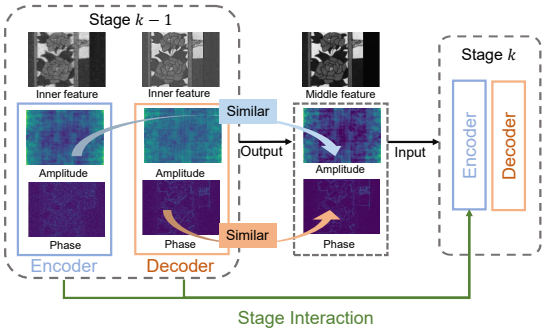

Figure 3: Visualization of inner features of deep unfolding networks in the frequency domain. The Amplitude/Phase image is extracted by the inner feature with Phase/Amplitude component set to a constant.

deep networks, resulting in inherent interpretability. The most well-known model-aided approaches include Plug-and-Play (PnP) [25] and Deep Unfolding [14, 32, 4]. Since PnP-based methods generally exploit a fixed pretrained denoiser and fail to learn a specific mapping for HSI, their reconstruction performance is hampered. Recently, deep unfolding-based methods [32, 14, 4] have achieved significant improvements in HSI reconstruction.

As deep unfolding framework unfolds the iterative optimization algorithms with neural networks, it typically includes a data module and a prior module. As illustrated in Figure 2, pixels in the 3D cube at different positions are compressed agnostically in the measurement, while the existing algorithm ignores such pixel-specific degradation in the data module. For the prior module, denoiser plays a critical role in the multi-stage optimization. Since HSIs are in 3D representation, it remains an issue for existing denoisiers to effective use of spatial-spectral information.

Besides, during the iterative recovery process, cross-stage fusion is necessary to prevent the loss of key information while gaining comprehensive features. We observe that the frequency information varies at different stages and depths (see Figure 3). The features from the previous encoder layer have clearer amplitude information, while the features from the previous encoder layer have clearer phase information. However, in the former works [32, 44, 4], the underlining differences of cross-stage features in the frequency domain are ignored, which incurs inferior performance of the multi-stage framework.

Inspired by the above findings, we propose a **P**ixel **A**daptive **D**eep **U**nfolding **T**ransformer (PADUT) framework for HSI reconstruction. **First**, we introduce the deep unfolding framework for the reconstruction process with the position-specific degradation information taken into consideration in the data module. **Second**, we propose a Non-local Spectral reconstruction Transformer for HSI to utilize the two-dimensional data of HSI in each stage. **Third**, we employ the frequency component analysis of HSI to fuse the features across iterative stages. With the observation that the encoder feature and the decoder feature have different emphases in the frequency domain, we propose Fast Fourier Transform Stage Fusion (FFT-SF) module, which leads to more comprehensive features for superb performance. The specific contributions of our work are:

- We propose a pixel adaptive deep unfolding transformer for HSI reconstruction. In the data module, we introduce pixel-level adaptive recovery at different locations. In the prior module, we propose a Non-local Spectral transformer for HSI processing.
- We introduce a novel frequency perspective for the cross-stage features in the iterative reconstruction process. Particularly, amplitude and phase representations are employed to establish interactions between different stages and depths.
- We carry out extensive experiments on both simulation scenes and real scenes to exhibit the effectiveness of our method for HSI reconstruction.

2. Related Work

In this section, we provide a brief overview of the recent advances in the field of HSI construction. Recent years have witnessed a paradigm shift from model-based methods to deep learning methods. Now, a progressive practice is to incorporate physical constraints with a data-driven network.

2.1. Model-based methods

Model-based HSI reconstruction methods [1, 37, 33, 43, 22, 42, 10] commonly optimize the objective function by separating the data fidelity term and regularization term. Twist [1] introduced a two-step Iterative shrinkage/thresholding algorithm for reconstruction with missing samples. The non-local similarity and low-rank regularization were utilized in [9] and [10]. The sparse representation [21, 33] and Gaussian mixture model [35] have also been widely studied. In [37], generalized alternating projection (GAP) method was proposed for HSI compressive sensing with utilized the total variation minimization. Although these methods produce satisfying results in the case of proper situation, they still face the problems of lacking generalization ability and computational efficiency.

2.2. Deep Learning-based methods

Inspired by the remarkable success of deep neural networks, deep learning-based methods [31, 8, 3, 20] have gained widespread utilization in the HSI reconstruction area. The pioneering methods SDA [28] and Reconnet [19] demonstrated the effectiveness of deep networks compared to traditional methods. Later, TSA-Net [24] introduced the spatial-spectral self-attention to sequentially reconstruct HSI. An external and internal learning method was introduced in [8] for utilizing the image-specific prior. To explore the power of Transformer for compressive sensing, MST [3] and CST [39] were proposed to capture the inner similarity of HSIs. Hu *et al*. [12] introduced the high-resolution dual-domain learning network (HDNet) to solve the spectral compressive imaging task. Though significant progress has been made, it is difficult for these brute-force methods to utilize the physical degradation characteristics.

2.3. Interpretable networks methods

Model-guided interpretable networks for HSI denoising have been actively explored in [32, 2, 34, 3, 44]. On the one hand, traditional handcrafted prior is employed in deep networks as a component [29]. On the other hand, conventional optimization algorithms can be represented by recurrent deep networks, allowing for the application of deep learning techniques to optimization problems. The plug-and-play methods [38, 45] plugged the deep denoisers into the optimization process. Deep unfolding HSI reconstruction methods [32] enjoyed the power of deep learning and known degradation process. For instance, GAP-Net unfolded the generalized alternating projection algorithm with a trained convolutional neural network. DAUHST [4] introduced a novel half-shuffle Transformer into the unfolding framework. Different from those methods that have limitations in exploring the pixel-specific degradation information and cross-stage features, our method introduces the pixel-adaptive recovery data module and utilizes the frequency information to guide the cross-stage fusion process.

3. Method

3.1. Problem Formulation

Based on the compressive theory [5], the CASSI system can capture a compressed measurement that includes information covering all bands. Figure 2 illustrates a basic pipeline of the coding procedure. Considering a spectral image $\mathbf{X}_\lambda \in \mathbb{R}^{M\times N}$ with λ as its wavelength, the captured HSI from real scenes is firstly modulated via a coded aperture $\mathbf{C}_\lambda \in \mathbb{R}^{M\times N}$. The temporary measurement $\mathbf{Y}_\lambda \in \mathbb{R}^{M\times N}$ is denoted as:

$$\mathbf{Y}_\lambda = \mathbf{C}_\lambda \odot \mathbf{X}_\lambda, \tag{1}$$

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

By shifting $\mathbf{Y}_\lambda$ along the horizontal direction according to the dispersive function d, the intermediate measurement $\mathbf{Y}'_\lambda \in \mathbb{R}^{\mathrm{M}\times(\mathrm{N}+\mathrm{B}-1)}$ is modulated to:

$$\mathbf{Y}'_\lambda(h,w) = \mathbf{G}'_\lambda(h, w+d(\lambda)), \tag{2}$$

where h and w denote the spatial coordinates. B is the number of bands in the desired 3D HSI. In the presence of noise $N \in \mathbb{R}^{M\times(N+B-1)\times B}$, the final measurement $\mathbf{Y} \in \mathbb{R}^{\mathrm{M}\times(\mathrm{N}+\mathrm{B}-1)\times \mathrm{B}}$ can be formulated as:

$$\mathbf{Y} = \sum_{\lambda=1}^{B} \mathbf{G}'_\lambda + \mathbf{N}. \tag{3}$$

For the connivance, with a shifted version of coded aperture $\mathbf{C}_\lambda$ as $\mathbf{\Phi}$, the overall imaging model is formulated as:

$$\mathbf{y} = \mathbf{\Phi}\mathbf{x} + \mathbf{n}. \tag{4}$$

CASSI system makes sacrifices the spatial information to obtain the spectral information. Consequently, the spatial intensity in the coded measurement $\mathbf{y}$ incorporates a combined representation of spatial and spectral information. This suggests that pixels in different locations in the HSI may have different levels of compression. This motivates us to improve in the optimization process for the pixel-specific reconstruction.

3.2. Revisting the Deep Unrolling Framework

Mathematically, the optimization of HSI reconstruction could be modeled as:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}||\mathbf{y} - \Phi\mathbf{x}||_2{}^2 + \eta J(\mathbf{x}), \tag{5}$$

where $J(\mathbf{x})$ denoted the regularizer term with parameter η.

The coding mask Φ reveals the spatial relation as well as the spectral relation between the coded measurement and desired 3D data. Recent deep unfolding works [32, 44, 4] have shown the great potential of combing Φ with deep networks. In the half quadratic splitting (HQS) algorithm, Eq. (5) is formulated into subproblems through the introduction of an auxiliary variable $\mathbf{z}$ as:

$$(\hat{\mathbf{x}}, \hat{\mathbf{z}}) = \arg\min_{\mathbf{x},\mathbf{z}} \frac{1}{2}||\mathbf{y} - \Phi\mathbf{x}||_2^2 + \eta J(\mathbf{z}) + \frac{\mu}{2}||\mathbf{z} - \mathbf{x}||_2^2, \tag{6}$$

where μ is the penalty parameter. HQS algorithm approximatively optimizes Eq. (6) through two iterative convergence subproblems:

$$\mathbf{x}^{\mathrm{k}+1} = \arg\min_{\mathbf{x}} ||\mathbf{y} - \Phi\mathbf{x}^{\mathbf{k}}||_2^2 + \mu||\mathbf{z}^{\mathrm{k}} - \mathbf{x}^{\mathrm{k}}||_2^2, \tag{7}$$

$$\mathbf{z}^{\mathrm{k}+1} = \arg\min_{\mathbf{z}} \frac{\mu}{2}||\mathbf{z}^{\mathbf{k}} - \mathbf{x}^{\mathrm{k}+1}||_2^2 + \eta J(\mathbf{z}^{\mathbf{k}}), \tag{8}$$

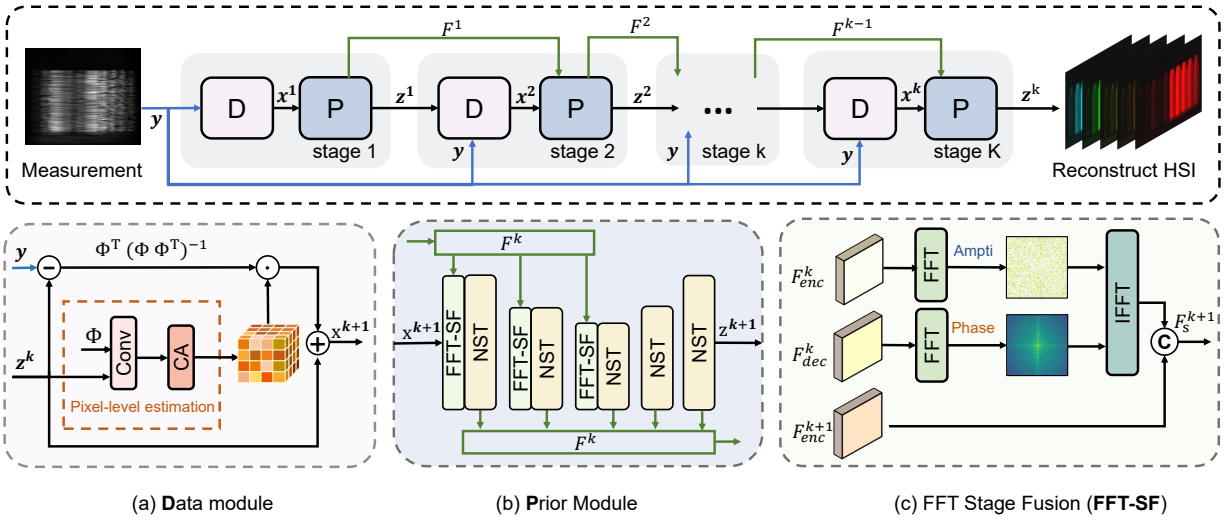

Figure 4: Illustration of our proposed Pixel Adaptive Deep Unfolding Transformer (PADUT) for HSI reconstruction. Top: the overall architecture that consists of K stages, each of which consists of a data module and a prior module. (a) Pixel-adaptive data module (b) Prior module (c) Stage fusion module

In short, a close-form solution of $\mathbf{x}$ is formulated by:

$$\mathbf{x}^{k+1} = (\mathbf{\Phi}^T\mathbf{\Phi} + \mu\mathbf{I})^{-1}(\mathbf{\Phi}\mathbf{y} + \mu\mathbf{z}^k) \tag{9}$$

$$= \mathbf{z}^k + \frac{1}{1+\mu}\mathbf{\Phi}^T(\mathbf{\Phi}\mathbf{\Phi}^T)^{-1}(\mathbf{y} - \mathbf{\Phi}\mathbf{z}^k). \tag{10}$$

In deep unrolling methods, $\mathbf{z}$ is often solved by:

$$\mathbf{z}^{k+1} = \mathrm{P}_{k+1}(\mathbf{x}^{k+1}), \tag{11}$$

where P_{k+1} refers to the deep network in the $k+1$ stage.

Usually, a deep unfolding framework consists of multiple stages specifically devised to reconstruct the underlying HSI cube $\mathbf{x}$ from coded measurement $\mathbf{y}$. Eq. (10) serves as a data module that introduces the physical characteristics into optimization. Meanwhile, deep characteristics are exploited in Eq. (11) and can be referred to the prior module.

As mentioned in the problem formulation, pixels in the HSI suffer varying degrees of information loss in the compressive sensing. Although the physical mask $\mathbf{\Phi}$ alleviates such a problem in the prior module, it often takes a fixed way of assistance. Moreover, in a real sensing system, there is often a gap between the mask and the real degradation.

3.3. Framework

Based on the aforementioned observation, we design a pixel-adaptive deep unfolding transformer for HSI reconstruction. Figure 4 illustrates the general framework of our proposed approach, which is composed of K stages to reconstruct a compressed HSI. In each stage, a data module is followed by a denoiser, which refers to the prior module. The data module aims to utilize the physical degradation information while the prior module is for optimization. Our denoiser is a U-shaped design. In the encoder, each layer contains a Fast Fourier Transformer stage fusion (FFT-SF) layer and a Non-local Spectral Transformer (NST) layer. The decoder is only composed of NST layers.

Pixel-Adaptive Prior Module. Observing from Eq. (10), $\frac{1}{1+\mu}$ plays an important role in the optimization of $\mathbf{x}$. For simplify, we use $\mathbf{F}_\sigma$ to represent $\frac{1}{1+\mu}$ as:

$$\mathbf{x}^{k+1} = \mathbf{z}^k + \mathbf{F}_\sigma\mathbf{\Phi}^T(\mathbf{\Phi}\mathbf{\Phi}^T)^{-1}(\mathbf{y} - \mathbf{\Phi}\mathbf{z}^k). \tag{12}$$

In the compressive sensing process, patterns in different positions and bands are markedly different due to the modulation. Due to the presence of instrument noise, the distribution of noise is also varied in the HSI cube. This difference persists throughout the recovery process. Considering the problem of inconsistent and agnostic degradation at different locations in the HSI, we design a pixel-adaptive data module for the deep unfolding framework.

The details of our pixel-adaptive prior module are illustrated in Figure 4 (a). Since the physical mask Φ establishes a relevance of the spatial and spectral dimensions, and $\mathbf{z}^k$ indicates the current input feature, we generate the 3D parameters $\mathbf{F}_\sigma$ via the convolution layer and Channel Attention (CA) [11] layer.

$$\mathbf{F}'_\sigma = \mathrm{Conv}(\mathrm{Concat}[\mathbf{z}^k, \mathbf{\Phi}]), \tag{13}$$

$$\mathbf{F}_\sigma = \mathrm{CA}(\mathbf{F}'_\sigma), \tag{14}$$

Then, the obtained 3D parameters $\mathbf{F}_\sigma$ are used to parameterize the 3D data in eq. (12) by the pixel-adaptive gradient descent step, achieving the pixel specific reconstruction.

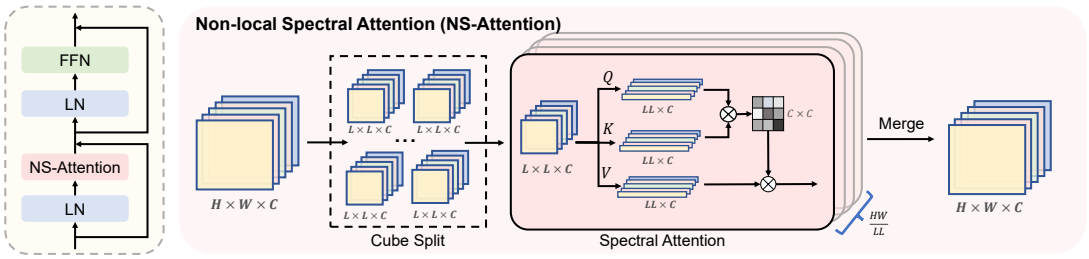

Figure 5: Illustration of our Non-local Spectral Transformer (NST) layer. The core modules of our block are: LayerNorm (LN), feed-forward network (FFN) and NS-Attention module.

Non-local Spectral Transformer. The spectral self-attention [3, 40] has shown a promising result in the image restoration area. However, it can hardly model the fine-grained similarity characteristic between pixels in both spatial and spectral dimensions. On the one hand, since spectral self-attention takes the pixels of the entire spectral dimension as the feature value to represent the spectral characteristic, local detail information can easily be lost. On the other hand, due to the coding property of the CASSI system, compressed information can often be found in adjacent areas. The spectral self-attention needs to better adapt to the 3D HSI cube and the coding system.

To make use of spatial-spectral information of HSI, we propose the Non-local Spectral Transformer (NST) for HSI restoration in the prior module. As shown in Figure 5, NST consists of Layer Normalization (LN), a Non-local Spectral Attention (NSA), and a Feed-Forward Network (FFN). The spatial shift operation is conducted between two NST to explore more than local features.

For the non-local spectral attention layer, we first split the entire feature of $\mathbf{x}_{\text{in}} \in \mathbb{R}^{\text{H}\times\text{W}\times\text{C}}$ into several cube patches as $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{\text{G}}\}$. Each cube is of size $L\times L\times C$. For each cube, we project $\mathbf{x}_i \in \mathbb{R}^{LL\times C}$ into query $\mathbf{Q}_i \in \mathbb{R}^{LL\times C}$, key $\mathbf{K}_i \in \mathcal{R}^{LL\times C}$, and value $\mathbf{Q}_i \in \mathbf{R}^{LL\times C}$ as

$$\mathbf{Q}_i = \mathbf{x}_i\mathbf{W}^{\text{Q}}, \mathbf{K}_i = \mathbf{x}_i\mathbf{W}^{\mathbf{K}}, \mathbf{V}_i = \mathbf{x}_i\mathbf{W}^{\mathbf{Q}}. \tag{15}$$

After the projection, the self-attention features for each cube are calculated as:

$$\text{Attention}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) = \mathbf{V}_i\text{Softmax}(\frac{\mathbf{K}_{\text{i}}{}^{\text{T}}\mathbf{Q}_{\text{i}}}{\beta}), \tag{16}$$

where the obtained attention map is of size $\mathbb{R}^{C\times C}$ for each spatial-spectral cube, capturing and consolidating the non-local information across the entire data volume. In the implementation, we adopt a similar approach of multi-head self-attention and partition the number of spectral bands into 'heads' and subsequently learn individual features.

Fast Fourier Transform Stage Fusion. Deep unfolding framework has shown the effectiveness of multi-stage learning via interpretable networks. Since the contextual information and detailed information varied at different stages, effectively employing the rich features could boost the performance of reconstruction [44, 27]. Moreover, inside each stage, the encoder-decoder denoiser leads to contextually different intermediate features due to the inherent trade-off between spatial and spectral information. How to interpolate cross-stage features and inner-stage features more effectively remains an ongoing challenge.

As shown in Figure 3, in the frequency domain, the phase component and amplitude component of recovery HSI in different stages correspond differently. In the encoder, the magnitude information is more prominent. In the later decoder, the phase information is more clear. According to this observation, we introduce the Fast Fourier Transform to the inter-stage connection to obtain a better reconstruction result from the frequency domain.

The details of our FFT-SF is shown in Figure 4 (c). We first transform the encoder and decoder feature from the former layer into Fourier domain. Then, a Fourier-based fusion is conducted to focus on the different frequency characteristics. Last, the frequency-enhanced feature is used to enhance the feature of next stage.

To model the frequency feature of $\mathbf{x}$ with a shape of $\mathbb{R}^{H\times W\times C}$, we leverage the Fourier transform $\mathcal{F}$ to convert it into Fourier domain, which is formulated as $\mathcal{F}(\mathbf{x})$:

$$\mathcal{F}(\mathbf{x}) = \mathbf{x}(u, v) \tag{17}$$

$$= \frac{1}{\sqrt{HW}}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\mathbf{x}(h, w)e^{-j2\pi(\frac{h}{H}u+\frac{w}{W}v)}, \tag{18}$$

where u and v stand for coordinates in frequency domain.

To analyze and utilize the frequency characteristics of HSIs, we decompose the complex component $\mathbf{x}(u, v)$ into amplitude $A(\mathbf{x})$ and phase $P(\mathbf{x})$. The amplitude component provides insight into the intensity of pixels, whereas phase component is critical for conveying positional information. Following [13], the mathematical formulation is given by:

$$A(\mathbf{x}(u, v) = \sqrt{R^2(\mathbf{x}(u, v)) + I^2(\mathbf{x}(u, v))}, \tag{19}$$

$$P(\mathbf{x}(u, v)) = \arctan[\frac{I(\mathbf{x}(u, v)}{R(\mathbf{x}(u, v))}], \tag{20}$$

	Params	GFLOPs	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	Avg
TwIST [1]	-	-	25.16 0.700	23.02 0.604	21.40 0.711	30.19 0.851	21.41 0.635	20.95 0.644	22.20 0.643	21.82 0.650	22.42 0.690	22.67 0.569	23.12 0.669
GAP-TV [37]	-	-	26.82 0.754	22.89 0.610	26.31 0.802	30.65 0.852	23.64 0.703	21.85 0.663	23.76 0.688	21.98 0.655	22.63 0.682	23.1 0.584	24.36 0.669
DeSCI [22]	-	-	27.13 0.748	23.04 0.620	26.62 0.818	34.96 0.897	23.94 0.706	22.38 0.683	24.45 0.743	22.03 0.673	24.56 0.732	23.59 0.587	25.27 0.721
λ-Net [26]	62.64M	117.98	30.10 0.849	28.49 0.805	27.73 0.870	37.01 0.934	26.19 0.817	28.64 0.853	26.47 0.806	26.09 0.831	27.50 0.826	27.13 0.816	28.53 0.841
TSA-Net [24]	44.25M	110.06	32.03 0.892	31.00 0.858	32.25 0.915	39.19 0.953	29.39 0.884	31.44 0.908	30.32 0.878	29.35 0.888	30.01 0.890	29.59 0.874	31.46 0.894
DGSMP [14]	3.76M	646.65	33.26 0.915	32.09 0.898	33.06 0.925	40.54 0.964	28.86 0.882	33.08 0.937	30.74 0.886	31.55 0.923	31 .66 0.911	31 .44 0.925	32.63 0.917
GAP-Net [23]	4.27M	78.58	33.74 0.911	33.26 0.900	34.28 0.929	41.03 0.967	31.44 0.919	32.40 0.925	32.27 0.902	30.46 0.905	33.51 0.915	30.24 0.895	33.26 0.917
HDNet [12]	2.37M	154.76	35.14 0.935	35.67 0.940	36.03 0.943	42.30 0.969	32.69 0.946	34.46 0.952	33.67 0.926	32.48 0.941	34.89 0.942	32.38 0.937	34.97 0.943
MST-L [3]	2.03M	28.15	35.40 0.941	35.87 0.944	36.51 0.953	42.27 0.973	32.77 0.947	34.80 0.955	33.66 0.925	32.67 0.948	35.39 0.949	32.50 0.941	35.18 0.948
CST-L [39]	3.00M	40.01	35.96 0.949	36.84 0.955	38.16 0.962	42.44 0.975	33.25 0.955	35.72 0.963	34.86 0.944	34.34 0.961	36.51 0.957	33.09 0.945	36.12 0.957
DAUHST-L [4]	6.15M	79.50	37.25 0.958	39.02 0.967	41.05 0.971	46.15 0.983	35.80 0.969	37.08 0.970	37.57 0.963	35.10 0.966	40.02 0.970	**34.59** 0.956	38.36 0.967
PADUT-3stg	1.35M	22.91	36.25 0.951	37.92 0.963	39.63 0.970	44.55 0.985	34.59 0.964	35.58 0.965	35.69 0.950	33.76 0.960	38.26 0.963	33.24 0.947	36.95 0.962
PADUT-5stg	2.24M	37.90	36.68 0.955	38.74 0.969	41.37 0.975	45.79 0.988	35.13 0.967	36.37 0.969	36.52 0.959	34.40 0.967	39.57 0.971	33.78 0.955	37.84 0.967
PADUT-7stg	3.14M	52.90	37.34 0.961	39.74 0.974	41.92 0.976	**47.01** 0.990	35.70 0.971	36.73 0.972	37.01 0.960	34.68 0.970	39.51 0.972	34.43 0.961	38.41 0.971
PADUT-12stg	5.38M	90.46	**37.36** **0.962**	**40.43** **0.978**	**42.38** **0.979**	46.62 **0.990**	**36.26** **0.974**	**37.27** **0.974**	**37.83** **0.966**	**35.33** **0.974**	**40.86** **0.978**	34.55 **0.963**	**38.89** **0.974**

Table 1: Results on 10 simulated scenes from CAVE dataset (S1∼S10). The best results are in bold.

where $R(\mathbf{x})$ and $I(\mathbf{x})$ present the real and imaginary parts.

For the (k+1)-th stage, the feature from the former stage denotes as F^k_{enc} and F^k_{dec}, feature from the current encoder layer as F^{k+1}_{enc}. Based on the observation that amplitude information and phase information are expressed differently in encoder and decoder, FFT stage fusion is expressed as:

$$F_s^{k+1\prime} = \mathcal{F}^{-1}(A(F^k_{enc}), P(F^k_{dec})), \tag{21}$$

$$F_s^{k+1} = \mathrm{Conv}(\mathrm{Concat}(F_s^{k+1\prime}, F^{k+1}_{enc})]), \tag{22}$$

where $\mathcal{F}^{-1}$ represents the inverse Fourier transform.

4. Experiments

In this section, we initially introduce the experimental setup and implementation specifics. Subsequently, we assess the performance of our proposed method on both synthetic data and real-world HSI datasets. Lastly, we conduct an ablation study to showcase the efficacy of our approach.

Datasets. For the simulated experiment, we utilized two widely used HSI datasets, namely CAVE and KAIST, for training and testing. The KAIST dataset consists of 30 HSIs with a spatial resolution at 2704×3376 and a spectral dimension of 31. The CAVE dataset comprises 32 HSIs of size 512× 512 × 31. In accordance with the settings in TSA-Net [24], we adopt CAVE dataset as the training set and 10 scenes from KAIST dataset as testing set. The patch size of each HSI is 256×256×28.

For the real data experiment, five real HSIs collected in TSA-Net [24] are used for evaluation. Each testing sample is of size 660×660×28. Following [3], training samples are extracted from the CAVE dataset and KAIST dataset with the patch size of 660×660×28.

Implementation. Our PADUT is implemented with Pytorch and trained with Adam [18] optimizer for 300 epochs. During training, the learning rate is 4×10^{-4} using the cosine annealing scheduler. The batch size is set to 5.

Competing Methods. We compare our method with three classic model-based spectral reconstruction methods (TwIST [1], GAP-TV [37] and DeSCI [22]), five end-to-end methods (Lambda-Net [26], TSA-Net [24], MST [3], HDNet [12] and CST[39]) and three deep unfolding methods (GAP-Net [23], DGSMP [14], and DAUHST [4]).

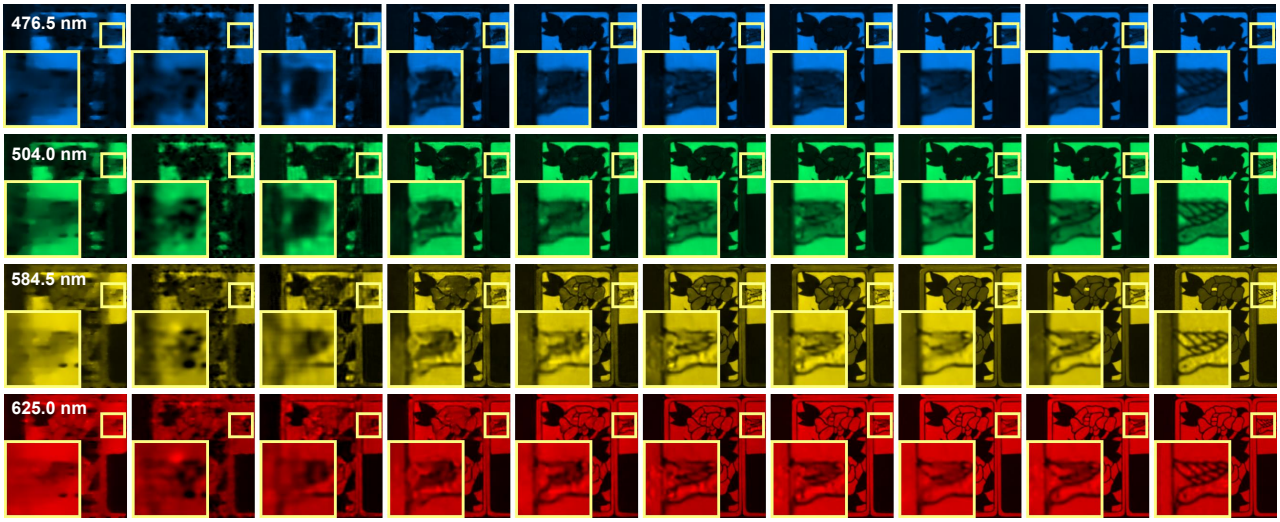

Figure 6: Visual comparisons on the KAIST dataset of scene 07 with 4 spectral channels.

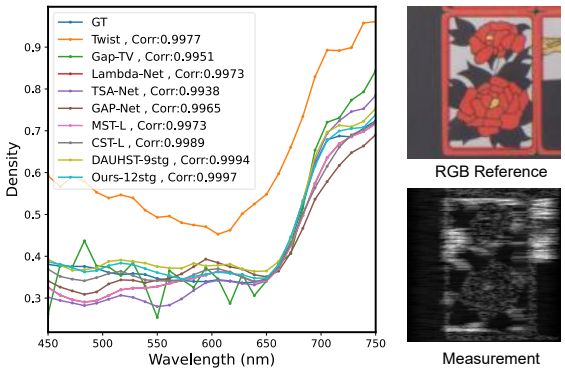

Figure 7: Spectral density curve on the simulation dataset of scene 07 with its corresponding RGB image and compressed measurement.

Evaluation Metrics. The reconstruction quality is evaluated using peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM).

4.1. Simulation Results

Numerical Results. The results from 10 simulated scenes are represented in Table 1. From the numerical results, we can observe that our method achieves the best result in almost all scenes and metrics, verifying the effectiveness of our method. Compared to end-to-end methods MST-L and CST-L, our PADUT-3stg achieves an improvement of 0.8 dB in PSNR. Our larger version PADUT-7stg outperforms DAUHST-L with a cost of 66.5% (52.9/79.5) GFLOPs. This highlights the benefit of leveraging the intrinsic characteristics of the CASSI system in pixel-level. Moreover, our method performs particularly well on the metric of SSIM. Figure 1 reports the SSIM-GFLOPs comparisons of our method and recent HSI reconstruction methods.

Baseline	PA	FFT-SF	NST	Params	GFLOPs	PSNR	SSIM
✓				1.30M	20.31	36.19	0.959
✓	✓			1.33M	22.41	36.37	0.961
✓	✓	✓		1.35M	22.91	36.84	0.962
✓	✓	✓	✓	1.35M	22.91	**36.95**	**0.962**

Table 2: Break-down ablation study on individual components of the proposed method.

Visual Results. For better vision, following [24], we show the visual results in RGB format with CIE color as the mapping function. Figure 6 demonstrates that our method excels in preserving clearer spatial details, particularly in the zoomed area. Figure 7 illustrates the corresponding spectral curves of competing methods as well as our method. The curve of our method is closest to the GroudTruth, indicating the spectral fidelity of our method.

4.2. Real Results

The visual results of a real-scene reconstructed HSI are shown in Figure 8. While most methods fail to reconstruct the detailed texture details, our method can successfully recover clear details. Compared to the deep unfolding-based approach DAUHST, we introduce the pixel-level reconstruction data module and FFT-based stage fusion, thus achieving better reconstruction results.

4.3. Ablation Study

To verify the effectiveness of our proposed structure, we conduct the ablation study with Ours-3stg method. All the evaluations are conducted on the simulated datasets.

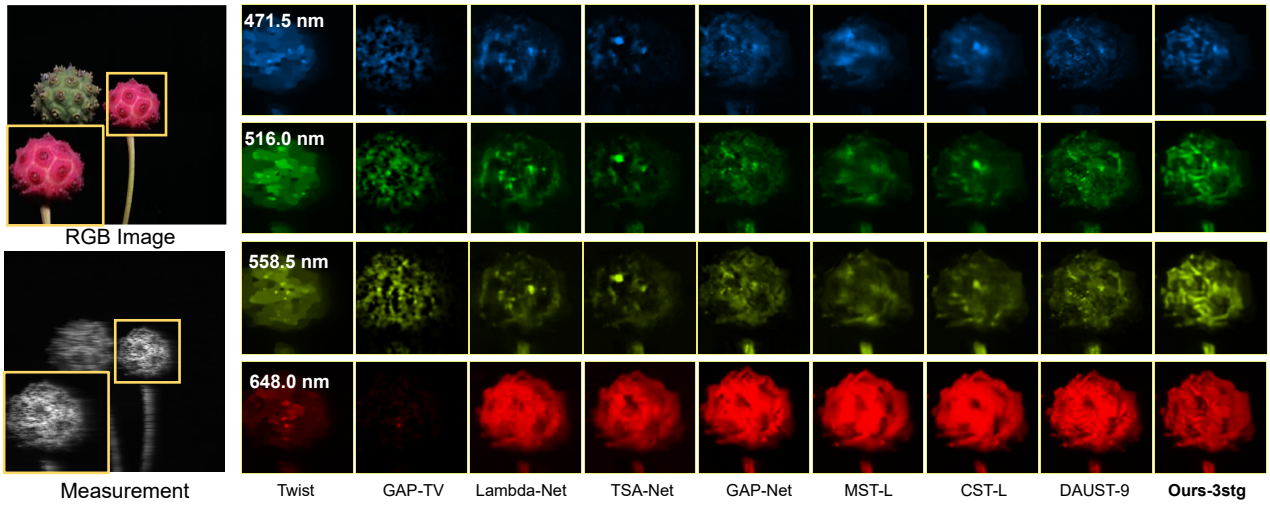

Figure 8: Reconstructed real HSI comparison on scene 4 from the real dataset.

	Baseline	Concat	ISFF	FFT-SF (Ours)
Params	1.33M	1.47M	1.35M	1.35M
GFLOPs	22.41	26.71	23.00	22.91
PSNR	36.41	36.60	36.53	**36.95**
SSIM	0.954	0.959	0.959	**0.962**

Table 3: Ablation study of different stage fusion.

	Params	GFLOPs	PSNR	SSIM
Sharing	0.46M	22.91	35.65	0.953
No-Sharing	1.35M	22.91	**36.95**	**0.962**

Table 4: Ablation study of the independent network parameters and parameter sharing.

Break-down Ablation. Here, we present the break-down ablation experiments on each component of our proposed framework. The Baseline is derived by removing the FFT-SF and pixel adaptive (PA) estimation module. Denoisier is replaced by Restormer [40]. The experimental results are shown in Table 2. It shows that the use of pixel-(PA) module leads to improved performance of 0.18 dB (from 36.19 dB to 36.37 dB). When we take the FFT-SF out, there is a drop of PSNR with 0.37 dB. Removing all the components, the performance degrades from 36.95 dB to 36.19 dB.

Ablation Study of Stage Interaction. In Table. 3, we evaluate the efficacy of the proposed FFT-SF. We compare FFT-SF with two other strategies of fusion, *i.e.*, directly Concat and ISFF [27]. One could see that the stage interaction is important to deep unfolding framework, since it prevents critical information loss. And our proposed FFT-SF obtains the best results. Since FFT-SF conducts the fusion in the frequency domain, it can restore better high-frequency details as shown in Figure 9.

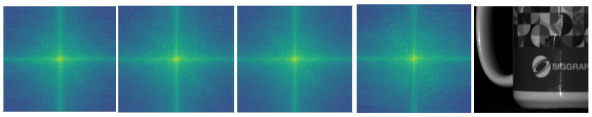

Figure 9: Visual comparison in the frequency domain when employing different fusion strategies.

Parameters Sharing. To further demonstrate our method, we conduct the ablation study on the parameter-sharing network. Specifically, the modules in different stages shares the same parameters. The results are shown in Table 4. Since our method employs the stage-targeted recovery by the pixel-adaptive data module, performance degrades when the weights of data module and denoisiers are shared.

5. Conclusion

In this paper, we propose a pixel adaptive deep unfolding transformer for HSI reconstruction. Our method aims to tackle the issues in existing deep unfolding works. In the data module, we employ the pixel-adaptive recovery, focusing on the imbalanced and agnostic degradation in CASSI. In the prior module, we introduce the Non-local Spectral Transformer to restore the HSI to emphasize the 3D characteristics. Moreover, inspired by the diverse expression of features in different stages and depths, the stage interaction is improved by the interaction in frequency domain. Experimental results reveal that our method surpasses the performance of the state-of-the-art HSI reconstruction methods.

Acknowledgments This work was supported by the National Natural Science Foundation of China (62171038, 61931008, 62006023, and 62088101), and the Fundamental Research Funds for the Central Universities.

References

[1] José M Bioucas-Dias and Mário AT Figueiredo. A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration. *TIP*, 16(12):2992–3004, 2007. 2, 6

[2] Théo Bodrito, Alexandre Zouaoui, Jocelyn Chanussot, and Julien Mairal. A trainable spectral-spatial sparse coding model for hyperspectral image restoration. volume 34, pages 5430–5442, 2021. 3

[3] Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *CVPR*, pages 17502–17511, 2022. 1, 3, 5, 6

[4] Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc Van Gool. Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging. 2022. 1, 2, 3, 6

[5] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006. 3

[6] Duncan T Eason and Mark Andrews. Total variation regularization via continuation to recover compressed hyperspectral images. *TIP*, 24(1):284–293, 2014. 1

[7] Baowei Fei. Hyperspectral imaging in medical applications. In *Data Handling in Science and Technology*, volume 32, pages 523–565. Elsevier, 2019. 1

[8] Ying Fu, Tao Zhang, Lizhi Wang, and Hua Huang. Coded hyperspectral image reconstruction using deep external and internal learning. *TPAMI*, 44(7):3404–3420, 2021. 3

[9] Ying Fu, Yinqiang Zheng, Imari Sato, and Yoichi Sato. Exploiting spectral-spatial correlation for coded hyperspectral image restoration. In *CVPR*, pages 3727–3736, 2016. 1, 2

[10] Wei He, Quanming Yao, Chao Li, Naoto Yokoya, Qibin Zhao, Hongyan Zhang, and Liangpei Zhang. Non-local meets global: An iterative paradigm for hyperspectral image restoration. *TPAMI*, 44(4):2089–2107, 2020. 2

[11] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 4

[12] Xiaowan Hu, Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Hdnet: High-resolution dual-domain learning for spectral compressive imaging. In *CVPR*, pages 17542–17551, 2022. 3, 6

[13] Jie Huang, Yajing Liu, Feng Zhao, Keyu Yan, Jinghao Zhang, Yukun Huang, Man Zhou, and Zhiwei Xiong. Deep fourier-based exposure correction network with spatial-frequency interaction. In *ECCV*, pages 163–180. Springer, 2022. 5

[14] Tao Huang, Weisheng Dong, Xin Yuan, Jinjian Wu, and Guangming Shi. Deep gaussian scale mixture prior for spectral compressive imaging. In *CVPR*, pages 16216–16225, 2021. 2, 6

[15] Sen Jia and Yuntao Qian. Constrained nonnegative matrix factorization for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 47(1):161–173, 2008. 1

[16] Nirmal Keshava. Distance metrics and band selection in hyperspectral processing with applications to material identification and spectral libraries. *IEEE Transactions on Geoscience and remote sensing*, 42(7):1552–1565, 2004. 1

[17] Muhammad Jaleed Khan, Hamid Saeed Khan, Adeel Yousaf, Khurram Khurshid, and Asad Abbas. Modern trends in hyperspectral image analysis: A review. *Ieee Access*, 6:14118–14129, 2018. 1

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[19] Kuldeep Kulkarni, Suhas Lohit, Pavan Turaga, Ronan Kerviche, and Amit Ashok. Reconnet: Non-iterative reconstruction of images from compressively sensed measurements. In *CVPR*, pages 449–458, 2016. 3

[20] Zeqiang Lai, Kaixuan Wei, and Ying Fu. Deep plug-and-play prior for hyperspectral image restoration. *Neurocomputing*, 481:281–293, 2022. 3

[21] Xing Lin, Yebin Liu, Jiamin Wu, and Qionghai Dai. Spatial-spectral encoded compressive hyperspectral imaging. *ACM Transactions on Graphics (TOG)*, 33(6):1–11, 2014. 2

[22] Yang Liu, Xin Yuan, Jinli Suo, David J Brady, and Qionghai Dai. Rank minimization for snapshot compressive imaging. *TPAMI*, 41(12):2990–3006, 2018. 1, 2, 6

[23] Ziyi Meng, Shirin Jalali, and Xin Yuan. Gap-net for snapshot compressive imaging. *arXiv preprint arXiv:2012.08364*, 2020. 6

[24] Ziyi Meng, Jiawei Ma, and Xin Yuan. End-to-end low cost compressive spectral imaging with spatial-spectral self-attention. In *ECCV*, pages 187–204. Springer, 2020. 1, 3, 6, 7

[25] Ziyi Meng, Zhenming Yu, Kun Xu, and Xin Yuan. Self-supervised neural networks for spectral snapshot compressive imaging. In *ICCV*, pages 2622–2631, 2021. 2

[26] Xin Miao, Xin Yuan, Yunchen Pu, and Vassilis Athitsos. l-net: Reconstruct hyperspectral images from a snapshot measurement. In *ICCV*, pages 4059–4069, 2019. 1, 6

[27] Chong Mou, Qian Wang, and Jian Zhang. Deep generalized unfolding networks for image restoration. In *CVPR*, pages 17399–17410, 2022. 5, 8

[28] Ali Mousavi, Ankit B Patel, and Richard G Baraniuk. A deep learning approach to structured signal recovery. In *2015 53rd annual allerton conference on communication, control, and computing (Allerton)*, pages 1336–1343. IEEE, 2015. 3

[29] Haiquan Qiu, Yao Wang, and Deyu Meng. Effective snapshot compressive-spectral imaging via deep denoising and total variation priors. In *CVPR*, pages 9127–9136, 2021. 3

[30] Ashwin Wagadarikar, Renu John, Rebecca Willett, and David Brady. Single disperser design for coded aperture snapshot spectral imaging. *Applied optics*, 47(10):B44–B51, 2008. 1

[31] Lizhi Wang, Chen Sun, Ying Fu, Min H Kim, and Hua Huang. Hyperspectral image reconstruction using a deep spatial-spectral prior. In *CVPR*, pages 8032–8041, 2019. 3

[32] Lizhi Wang, Chen Sun, Maoqing Zhang, Ying Fu, and Hua Huang. Dnu: Deep non-local unrolling for computational spectral imaging. In *CVPR*, pages 1661–1671, 2020. 2, 3

[33] Lizhi Wang, Zhiwei Xiong, Guangming Shi, Feng Wu, and Wenjun Zeng. Adaptive nonlocal sparse representation for

dual-camera compressive hyperspectral imaging. *TPAMI*, 39(10):2104–2111, 2016. 2

[34] Fengchao Xiong, Jun Zhou, Qinling Zhao, Jianfeng Lu, and Yuntao Qian. Mac-net: Model-aided nonlocal neural network for hyperspectral image denoising. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021. 1, 3

[35] Jianbo Yang, Xuejun Liao, Xin Yuan, Patrick Llull, David J Brady, Guillermo Sapiro, and Lawrence Carin. Compressive sensing by learning a gaussian mixture model from measurements. *TIP*, 24(1):106–119, 2014. 2

[36] Qiangqiang Yuan, Qiang Zhang, Jie Li, Huanfeng Shen, and Liangpei Zhang. Hyperspectral image denoising employing a spatial–spectral deep residual convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2):1205–1218, 2018. 1

[37] Xin Yuan. Generalized alternating projection based total variation minimization for compressive sensing. In *ICIP*, pages 2539–2543. IEEE, 2016. 1, 2, 6

[38] Xin Yuan, Yang Liu, Jinli Suo, and Qionghai Dai. Plug-and-play algorithms for large-scale snapshot compressive imaging. In *CVPR*, pages 1447–1457, 2020. 3

[39] Wulian Yun, Mengshi Qi, Chuanming Wang, Huiyuan Fu, and Huadong Ma. Coarse-to-fine video denoising with dual-stage spatial-channel transformer. *arXiv preprint arXiv:2205.00214*, 2022. 3, 6

[40] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5728–5739, 2022. 5, 8

[41] Shang Zhang, Yuhan Dong, Hongyan Fu, Shao-Lun Huang, and Lin Zhang. A spectral reconstruction algorithm of miniature spectrometer based on sparse optimization and dictionary learning. *Sensors*, 18(2):644, 2018. 1

[42] Shipeng Zhang, Hua Huang, and Ying Fu. Fast parallel implementation of dual-camera compressive hyperspectral imaging system. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(11):3404–3414, 2018. 2

[43] Shipeng Zhang, Lizhi Wang, Ying Fu, Xiaoming Zhong, and Hua Huang. Computational hyperspectral imaging based on dimension-discriminative low-rank tensor recovery. In *ICCV*, pages 10183–10192, 2019. 1, 2

[44] Xuanyu Zhang, Yongbing Zhang, Ruiqin Xiong, Qilin Sun, and Jian Zhang. Herosnet: Hyperspectral explicable reconstruction and optimal sampling deep network for snapshot compressive imaging. In *CVPR*, pages 17532–17541, 2022. 2, 3, 5

[45] Siming Zheng, Yang Liu, Ziyi Meng, Mu Qiao, Zhishen Tong, Xiaoyu Yang, Shensheng Han, and Xin Yuan. Deep plug-and-play priors for spectral snapshot compressive imaging. *Photonics Research*, 9(2):B18–B29, 2021. 3

[46] Liu Zhi, David Zhang, Jing-qi Yan, Qing-Li Li, and Qun-lin Tang. Classification of hyperspectral medical tongue images for tongue diagnosis. *Computerized Medical Imaging and Graphics*, 31(8):672–678, 2007. 1