DEGRADATION-AWARE UNFOLDING KNOWLEDGE-ASSIST TRANSFORMER FOR SPECTRAL COMPRESSIVE IMAGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Snapshot compressive spectral imaging offers the capability to effectively capture three-dimensional spatial-spectral images through a single-shot two-dimensional measurement, showcasing its significant potential for spectral data acquisition. However, the challenge of accurately reconstructing 3D spectral signals from 2D measurements persists, particularly when it comes to preserving fine-grained details like textures, which is caused by the lack of high-fidelity clean image information in the input compressed measurements. In this paper, we introduce a two-phase training strategy embedding high-quality knowledge prior in a deep unfolding framework, aiming at reconstructing high-fidelity spectral signals. Our experimental results on synthetic benchmarks and real-world datasets demonstrate the notably enhanced accuracy of our proposed method, both in spatial and spectral dimensions. Code and pre-trained models will be released.

1 INTRODUCTION

Hyperspectral images (HSIs) contain multiple spectral bands with more abundant spectral signatures than normal RGB images, widely applied in image classification (Maggiori et al., 2017; Li et al., 2019), object detection (Li et al., 2020; Rao et al., 2022), tracking (Van Nguyen et al., 2010; Uzkent et al., 2017), medical imaging (Lu & Fei, 2014; ul Rehman & Qureshi, 2021), remote sensing (Goetz et al., 1985; Lu et al., 2020), etc. To collect HSI, conventional imaging systems use spectrometers to scan the scenes along the spectral or spatial dimension, which is time-consuming and limited to static objects. Recently, snapshot compressive imaging (SCI) systems (Du et al., 2009; Llull et al., 2013; Cao et al., 2016; Luo et al., 2023) have obtained much attention to capture HSIs, among which the coded aperture snapshot spectral imaging (CASSI) Wagadarikar et al. (2008b); Meng et al. (2020c) stands out with impressive performance and efficiency. CASSI modulates the HSI signal at

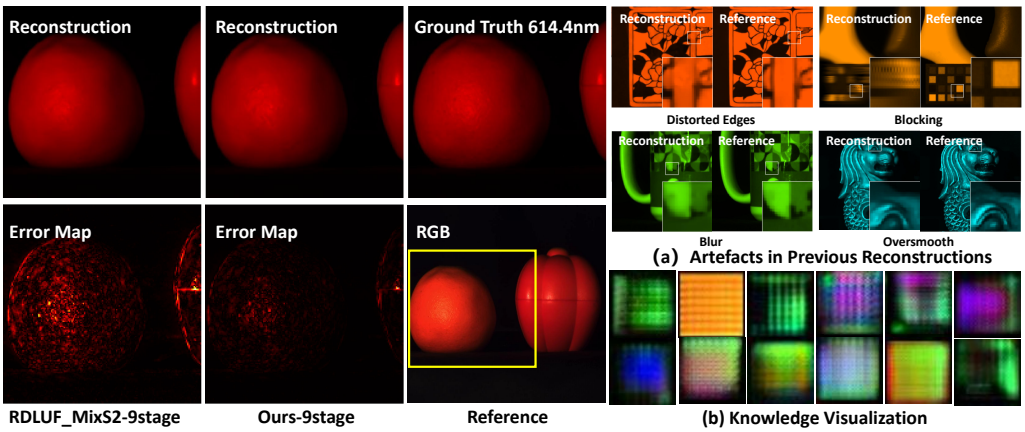

Figure 1: Comparisons with the previous reconstruction methods, and the visualization of learned knowledge. The error map shows the absolute error between the reconstructed spectral image and the ground truth, where brighter areas indicate larger error values. The zoomed parts in subfigures in (a) show several typical artefacts we found in previous reconstruction methods, including distorted edges, blocking, blur and oversmooth. The visualization of the learned knowledge in (b) further illustrates the learned patterns facilitate image fidelity.

different wavelengths and mixes all modulated spectra to generate a 2D compressed measurement. Subsequently, reconstructing the 3D HSI cube from the 2D compressive measurements poses a fundamental challenge for the CASSI system.

From traditional model-based methods Bioucas-Dias & Figueiredo (2007); Beck & Teboulle (2009); Yuan (2016) to widely used learning-based methodsCharles et al. (2011); Miao et al. (2019); Meng et al. (2020b), CASSI reconstruction methods have been developed for years to recover high-quality HSI signals. Recent studies adopt deep unfolding Wang et al. (2022); Cai et al. (2022c); Meng et al. (2023); Dong et al. (2023) framework with a multi-stage network to map the measurement into the HSI cube. These deep unfolding networks are intuitively interpretable by explicitly characterizing the image priors and the imaging system. Besides, these methods also enjoy the power of learning-based methods (deep denoisers or prior networks) and thus have great potential. Furthermore, with the help of the representation capabilities on the long-range dependency of the Transformer Vaswani et al. (2017), the deep denoisers of unfolding networks can explore the spatial and spectral correlations Hu et al. (2022); Cai et al. (2022b); Dong et al. (2023).

Previous reconstruction networks aim to recover the clean image from compressed measurement, which encourages learning the local and non-local similarity in spatial and spectral dimensions from measurement. However, there remains an intrinsic issue implied in the ill-posed CASSI inverse problem, where the compressed measurement is severely degraded due to physical modulation, spectral compression, and unpredictable system noise. Thus, the severe degradation mismatch leads to insufficient authenticity and limited fidelity in the reconstructed images, such as distorted edges, blocky patterns, blurring, and other artifacts (as shown in Fig. 1).

To solve the problem, in this paper, we leverage high-fidelity knowledge, *i.e.* discrete prototypes learned from the uncompressed images, and explicit degradation information into the proposed deep unfolding framework for spectral compressive image reconstruction. Specifically, we train a lightweight counselor model by constructing a vector quantized variational autoencoder (VQ-VAE) to learn high-fidelity knowledge from the ground-truth image space. Then, we proposed a U-net bridge denoiser equipped with the mask-aware knowledge-assist attention (MKA) mechanism to incorporate multi-layer degradation information and high-fidelity prior knowledge in an elegant and principal manner. Notably, our unfolding framework can absorb external prior knowledge from the counselor model and the cascade degradation information to boost its reconstruction performance. The main contributions we have made in this paper can be summarized as follows:

- *i*) We propose a degradation-aware knowledge-assist deep unfolding framework using vector-quantized high-fidelity HSI knowledge to guide the CASSI reconstruction.
- *ii*) We propose a U-net bridge denoiser that integrates high-fidelity HSI knowledge and cascade degradation information to guide the CASSI reconstruction, which is learned with a VQ-VAE-based counselor network and a sophisticated mask cross encoder, respectively.
- *iii*) Extensive experiments on the synthetic benchmark and real dataset demonstrate the superior accuracy of our proposed method both in spatial and spectral dimensions.

2 Related Work

2.1 Hyperspectral Image Reconstruction

Traditional model-based methods Wang et al. (2017); Zhang et al. (2019); Wagadarikar et al. (2008a) are mainly based on hand-crafted image priors, such as total variation Yuan (2016), sparsity Kittle et al. (2010), low-rank Liu et al. (2019), etc. While these methods have demonstrated theoretical properties and interpretability, they require manual parameter tuning with low reconstruction speed and suffer from limited representation capacity and poor generalization ability. Recently, deep learning methods have been used to solve the inverse problem of spectral SCI. The first popular branch is Plug-and-play (PnP) algorithms Chan et al. (2017); Ryu et al. (2019); Yuan et al. (2020; 2021), which integrate pre-trained denoising networks into traditional model-based methods to solve the HSI reconstruction problem, but suffering from fixed nature of pre-trained denoiser and limited performance. The second branch of deep learning methods follows the End-to-end (E2E) training manner, which usually employ a powerful network, such as convolutional neural network (CNN) Cheng et al. (2022); Lu et al. (2020); Hu et al. (2022) and Transformers Cai et al. (2022b;a), to learn the mapping function from measurements to desired HSIs. However, they learn a brute-force mapping

ignoring the working principles of CASSI systems, thereby lacking interpretability and flexibility for various hardware systems. To overcome these issues, Deep unfolding methods Meng et al. (2020a); Ma et al. (2019); Cai et al. (2022c); Dong et al. (2023) take advantage of both model-based methods and deep learning-based methods, which transfer conventional iterative optimization algorithms into the multi-stage network. Each stage typically involves a linear projection and a single-stage network that learns the underlying denoiser prior. Deep unfolding methods offer intuitive interpretability by explicitly modeling image priors and the system. Moreover, they benefit from the power of deep learning, presenting significant potential that remains partially untapped.

2.2 Prototype Learning for SCI

The key idea of prototype learning for SCI is to exploit the non-local similarity as prototypes across images. In the SCI problem considered here, the non-local similarity can also be searched across different frames. In previous methods, sparse representation with learned dictionaries has demonstrated its superiority in SCI reconstruction, such as sparse dictionary (usually over-complete) learning Aharon et al. (2006), Bayesian dictionary learning Yuan et al. (2015), group-based sparse coding Liu et al. (2019). However, these methods usually require an iterative optimization to learn the dictionaries and sparse coding or pay a price for running time in patch matching, suffering from high computational costs. With the development of VQVAE Van Den Oord et al. (2017), the first method introduced a highly compressed codebook, is realizable to learn the representative prototypes with a vector-quantized Autoencoder model. Unlike the large hand-crafted dictionary Jo & Kim (2021), the learnable codebook automatically learns optimal elements and provides superior efficiency and expressiveness, circumventing the laborious dictionary design Zhou et al. (2022). Inspired by prototype learning, this paper investigates a discrete proxy space for SCI reconstruction. Different from existing methods, we exploit the discrete unpolluted iconic prototypes with a counselor model from the ground truth video as external knowledge prior and aim to guide the reconstruction. Such designs allow our method to take full advantage of the knowledge so that it does not solely depend on the degraded compressed information, tightly fitting the spectral compressive imaging task and significantly enhancing the robustness of face restoration.

3 Proposed Model

3.1 Problem Formulation

The CASSI system aims to modulate different wavelengths in the spectral data cube and then integrate them into a 2D imaging sensor. The mathematical model of the single-disperser CASSI (SD-CASSI) Wagadarikar et al. (2008a) sensing process is illustrated as follows. As shown in the Fig. 2 (a), the original HSI data, denoted as $\boldsymbol{X} \in \mathbb{R}^{W\times H\times B}$, is modulated by the fixed physical mask $\boldsymbol{M} \in \mathbb{R}^{W\times H}$, where W, H, and B denote the width, height, and the number of spectral channels, respectively. The coded HSI data cube is represented as $\mathbf{X}'(:,:,b) = \mathbf{X}(:,:,b) \odot \mathbf{M}, b = 1,2,\ldots,B$, where $\odot$ represents the element-wise multiplication. After light propagating through the disperser, each channel of $\mathbf{X}'$ is shifted along the $\boldsymbol{H}$-*axis*. The shifted data cube is denoted as $\mathbf{X}'' \in \mathbb{R}^{W\times \tilde{H}\times B}$, where $\tilde{H} = H + d_B$. Here, d_B represents the shifted distance of the $B-th$ wavelength. This process can be formulated as modulating the shifted version $\tilde{\mathbf{X}} \in \mathbb{R}^{W\times \tilde{H}\times B}$ with a shifted mask $\tilde{\mathbf{M}} \in \mathbb{R}^{W\times \tilde{H}\times B}$, where $\tilde{\boldsymbol{M}}(i,j,b) = \boldsymbol{M}(w,h+d_B)$. At last, the imaging sensor captures the shifted image into a 2D measurement $\mathbf{Y}$, calculated as

$$\mathbf{Y} = \sum_{b=1}^{B} \tilde{\mathbf{X}}(:,:,b) \odot \tilde{\mathbf{M}}(:,:,b) + \mathbf{N}, \tag{1}$$

where $\mathbf{N} \in \mathbb{R}^{W\times \tilde{H}}$ denotes the measurement noise. By vectorizing the spectral data cube and measurement, that is $\boldsymbol{x} = \mathtt{vec}(\tilde{\mathbf{X}}) \in \mathbb{R}^{W\tilde{H}B}$ and $\boldsymbol{y} = \mathtt{vec}(\mathbf{Y}) \in \mathbb{R}^{W\tilde{H}}$, this model can be formulated as

$$\boldsymbol{y} = \mathbf{A}\boldsymbol{x} + \boldsymbol{n}, \tag{2}$$

where $\mathbf{A} \in \mathbb{R}^{W\tilde{H}\times W\tilde{H}B}$ denotes the sensing matrix (coded aperture) which is a concatenation of diagonal matrices, that is $\mathbf{A} = [\mathbf{D}_1,\ldots,\mathbf{D}_B]$, where $\mathbf{D}_b = Diag(vec(\tilde{\mathbf{M}}(:,:,b)))$ is the diagonal matrix with $vec(\tilde{\mathbf{M}}(:,:,b))$ as the diagonal elements. The sensing matrix $\mathbf{A}$ is a sparse matrix and $\boldsymbol{A}\boldsymbol{A}^\top$ is a diagonal matrix Jalali & Yuan (2019). In this paper, we focus on the ill-posed HSI restoration problem, recovering the high-quality image $\boldsymbol{x}$ from the compressed measurement $\boldsymbol{y}$.

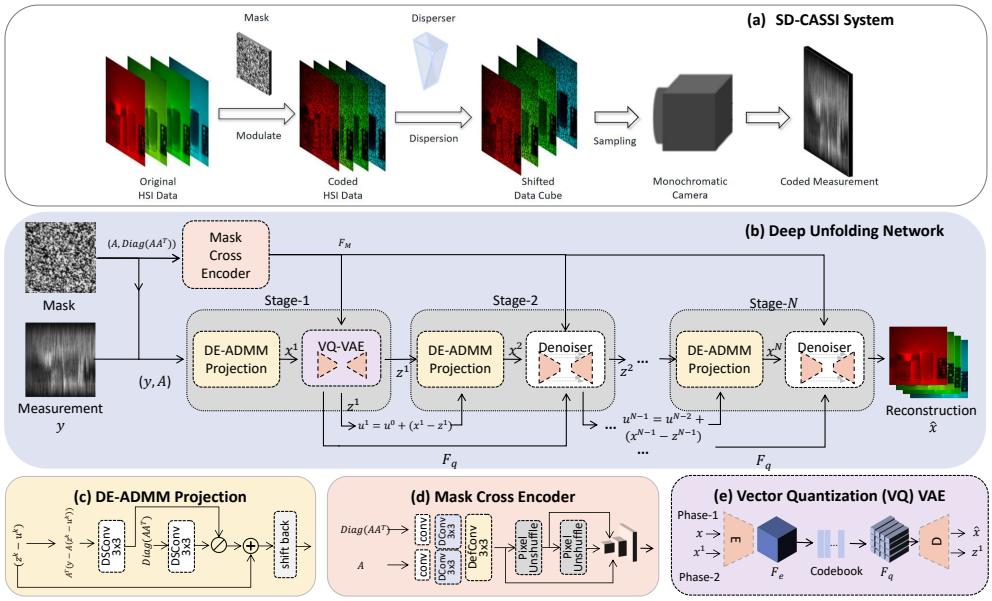

Figure 2: (a) The SD-CASSI forward process. HSI data cube is coded and compressed on a 2D sensor plane. (b) Our unfolding framework. (c) DE-ADMM projection. (d) Mask cross encoder. (e) Vector quantized Variational AutoEncoder (VQ-VAE). In phase 1, only VQ-VAE is trained in a self-supervised manner, both input and output are clean HSIs. In phase 2, VQ-VAE is embedded into the unfolding framework and updates its encoder parameter with other unfolding structures. Mask A and measurement y are input in this phase.

3.2 The Unfolding ADMM Framework

HSI reconstruction problem in Eq. 2 can be typically solved by convex optimization through the following objective:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda R(\boldsymbol{x}), \tag{3}$$

where λ is a noise-balancing factor. The first term guarantees that the solution $\hat{\boldsymbol{x}}$ meets the observation, and the second term $R(\boldsymbol{x})$ refers to the image regularization.

ADMM (Alternating Direction Method of Multipliers) Boyd et al. (2011) breaks down the original optimization problem into multiple sub-problems and updates them alternately, leading to excellent convergence, error reduction, reduced computational and storage complexity. Here, we choose ADMM as our optimization framework Zheng et al. (2021). Therefore, the problem in Eq. 3.2 can be written as:

$$\boldsymbol{x}^{k+1} = \arg\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \frac{\rho}{2}\left\|\boldsymbol{x} - \left(\boldsymbol{z}^k - \boldsymbol{u}^k\right)\right\|_2^2, \tag{4}$$

$$\boldsymbol{z}^{k+1} = \arg\min_{\boldsymbol{z}} \lambda R(\boldsymbol{z}) + \frac{\rho}{2}\left\|\boldsymbol{z} - \left(\boldsymbol{x}^{k+1} + \boldsymbol{u}^k\right)\right\|_2^2, \tag{5}$$

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + \left(\boldsymbol{x}^{k+1} - \boldsymbol{z}^{k+1}\right), \tag{6}$$

where $\boldsymbol{z}$ is an auxiliary variable, $\boldsymbol{u}$ is the multiplier, ρ is a penalty factor, and k is the index of iterations. Recalling the proximal operator Parikh et al. (2014), defined as $\text{prox}_g(\boldsymbol{v}) = \arg\min g(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{v}\|_2^2$, Equ. 5 is the Euclidean projection with a closed-form solution, i.e., $\boldsymbol{x}^{k+1} = \left(\boldsymbol{A}^\top\boldsymbol{A} + \rho\boldsymbol{I}\right)^{-1} \cdot \left[\boldsymbol{A}^\top\boldsymbol{y} + \rho\left(\boldsymbol{z}^k - \boldsymbol{u}^k\right)\right]$. Equ. 6 can be viewed as a denoiser $\mathscr{D}$. Furthermore, recalling that $\boldsymbol{A}\boldsymbol{A}^\top$ is a diagonal matrix for image-plane coding, $\left(\boldsymbol{A}^\top\boldsymbol{A} + \rho\boldsymbol{I}\right)^{-1}$ can be calculated efficiently using the matrix inversion lemma (Woodbury matrix identity) as $\left(\boldsymbol{A}^\top\boldsymbol{A} + \rho\boldsymbol{I}\right)^{-1} = \rho^{-1}\boldsymbol{I} - \rho^{-1}\boldsymbol{A}^\top\left(\boldsymbol{I} + \rho\boldsymbol{A}\boldsymbol{A}^\top\right)^{-1}\boldsymbol{A}\rho^{-1}$. Then the Euclidean projection can be simplified and the

final solution is

$$\boldsymbol{x}^{k+1} = \left(\boldsymbol{z}^k - \boldsymbol{u}^k\right) + \boldsymbol{A}^\top \left[\boldsymbol{y} - \boldsymbol{A}\left(\boldsymbol{z}^k - \boldsymbol{u}^k\right)\right] \oslash \left[\mathrm{Diag}\left(\boldsymbol{A}\boldsymbol{A}^\top\right) + \rho\right], \tag{7}$$

$$\boldsymbol{z}^{k+1} = \mathscr{D}_k\left(\boldsymbol{x}^{k+1} + \boldsymbol{u}^k\right), \tag{8}$$

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + \left(\boldsymbol{x}^{k+1} - \boldsymbol{z}^{k+1}\right), \tag{9}$$

where $\mathrm{Diag}(\cdot)$ extracts the diagonal elements of the ensued matrix, $\oslash$ denotes the Hadamard division, and $\mathscr{D}_k$ is the denoiser of the $k-th$ stage. Here, the noise penalty factor ρ is tuned to match the measurement noise. Considering the projection step Equ. 7, assisted by deep learning, we can correct this step as follows:

$$\boldsymbol{x}^{k+1} = \boldsymbol{z}_{\boldsymbol{u}}^k + DE(\boldsymbol{A}^\top(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_{\boldsymbol{u}}^k))DE\left(\mathrm{Diag}\left(\boldsymbol{A}\boldsymbol{A}^\top\right)\right), \tag{10}$$

where $\boldsymbol{z}_{\boldsymbol{u}}^k = \left(\boldsymbol{z}^k - \boldsymbol{u}^k\right)$, and $DE(\cdot)$ denotes the deep learning enhancement in ADMM, implemented by a neural network consisting of a depthwise separable convolution and a GELU Hendrycks & Gimpel (2016) operation. This enhancement aims to use deep linear and non-linear operations, as illustrated in Fig. 2 (c), to further correct the gradient descent direction in ADMM and estimate the influence of noise in the diagonal matrix. Thus, we name this step Deep Enhanced ADMM (DE-ADMM) Euclidean Projection. The overall unfolding framework is shown in Fig. 2 (b), mask A and measurement y are inputs of the network, following the Equ. 7,8 and 9, we acquire the first stage output $\boldsymbol{z}^1$ and residue $\boldsymbol{u}^1$. The first stage denoiser is a pre-trained VQ-VAE with some fixed modules. The intermediate variable $\mathbf{F}_q$ will be used as prior knowledge in the later stages. Other stage denoisers are U-shape networks, receiving cross space mask features and prior knowledge. Therefore, we will next introduce how to learn the cross-space mask features and high-fidelity knowledge priors, and then propose a novel scheme to integrate them to facilitate denoising and reconstruction.

3.3 Mask-aware High-fidelity knowledge Learning

3.3.1 Mask Cross Encoder

The mask is a crucial component in SCI reconstruction as it provides the degradation details in SCI measurement. However, establishing the relationship between the degradation information and the mask in compressed measurements can be challenging. To address this, we propose a Mask Cross Encoder (MCE) that extracts the Cross-Space Mask Feature from two Euclidean spaces: the observation and compressive domains. As shown in Fig. 2 (d), the mask ($\mathbf{A}$) and compressed mask ($Diag(\mathbf{A}\mathbf{A}^\top)$) are fed into a dual-path network respectively, and then two kinds of information $F_{\mathbf{A}\mathbf{A}^\top} = Conv(Sigmoid(Diag(\mathbf{A}\mathbf{A}^\top))$ and $F_{\mathbf{A}} = Conv(Sigmoid(\mathbf{A}))$ are concatenated and fused as:

$$F_{\text{fusion}} \in \mathbb{R}^{H\times W\times C} = DefConv(DConv(Concat(F_{\mathbf{A}}, F_{\mathbf{A}\mathbf{A}^\top})))$$

where $DConv$ refers to the dilated convolution network, which effectively increases the receptive field while minimizing the number of parameters. $DefConv$ denotes the deformable convolution, which learns offsets based on object shapes within images, allowing for the extraction of more intricate information compared to vanilla convolution. In this way, the mask-aware features across two Euclidean spaces are fused and aligned to guide the SCI reconstruction.

Then we use the pixel unshuffle (PU) operations to obtain the multi-level mask-aware representations by zooming feature size. Each PU operation rearranges elements, transforming a tensor of shape $H\times W\times C$ to $\frac{H}{2}\times\frac{W}{2}\times 4C$. As shown in Fig. 2 (d), two PU layers are employed, and the outputs contain three multi-scale mask-aware features, stated as:

$$F_M^i = \begin{cases} F_{\text{fusion}} & \text{if } i=1 \\ PU(F_M^{i-1}) & \text{if } i>1 \end{cases}, F_M^i \in \mathbb{R}^{\frac{H}{2^{i-1}}\times\frac{W}{2^{i-1}}\times 4^{i-1}C}, i=1,2,3.$$

To this end, MCE encodes mask features into hierarchical representations and will facilitate SCI reconstruction by integrating the degradation information.

3.3.2 Vector Quantization for High-fidelity Knowledge Learning

In previous SCI reconstruction methods, the input typically consists of compressed measurements and masks that lack local textures and details. This limited information results in the lack of high-fidelity natural image prior knowledge for the reconstruction. Thus, we aim to solve this problem using discrete representation to learn high-fidelity knowledge from the uncompressed domain.

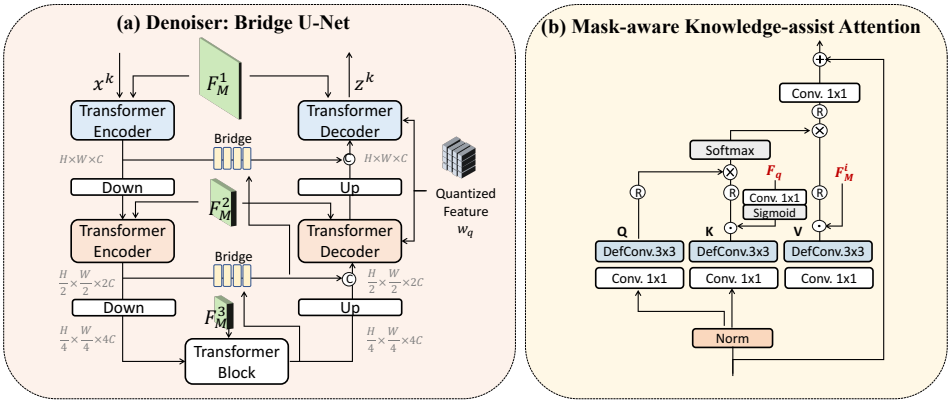

Figure 3: (a) The bridge U-Net. Each bridge module is composed of several convolution layers. The mask features obtained by the mask cross encoder will be sent into each encoder and decoder for cross-space reconstruction. (b) The Mask-aware Knowledge-assist Attention module (MKA), where the mask features F_M^i and prior knowledge F_q are integrated through cross attention.

Specifically, we design a two-phase vector quantized variational autoencoder to learn and integrate the discrete unpolluted knowledge.

Phase 1: We aim to learn the prior high-fidelity knowledge by a lightweight counselor model in a self-supervised manner, which encodes clean HSI data $\boldsymbol{x}$ into $\mathbf{F}_e$ by the encoder E, then reconstructs $\hat{\boldsymbol{x}}$ from $\mathbf{F}_q$ by the decoder D. Specifically, as shown in Fig. 2 (e), we first encode the $\boldsymbol{x} \in \mathbb{R}^{W\times H\times B}$ into latent feature $\mathbf{F}_e \in \mathbb{R}^{m\times n\times d}$ output by the ADMM projection. Then, we quantize the continual feature $\mathbf{F}_e$ as $\mathbf{F}_q \in \mathbb{R}^{m\times n\times d}$ by replacing the features of each position with its nearest prototypes in the codebook $\mathscr{C} = \left\{c_k \in \mathbb{R}^d\right\}_{k=0}^{N}$ along spatial dimension, formulated as:

$$\mathbf{F}_q^{(i,j)} = \underset{c_k\in\mathscr{C}}{\arg\min} \left\| \mathbf{F}_e^{(i,j)} - c_k \right\|_2 .$$

Then the quantized feature $\mathbf{F}_q$ is decoded into reconstruction $\hat{\boldsymbol{x}}$ through decoder D. The training objective for optimizing VQ-VAE and codebook can be formulated as:

$$\mathscr{L}_{\text{codebook}} = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_1 + \|\operatorname{sg}(\mathbf{F}_e) - \mathbf{F}_q\|_2^2 + \beta \left\|\mathbf{F}_e - \operatorname{sg}(\mathbf{F}_q)\right\|_2^2,$$

where $\operatorname{sg}(\cdot)$ stands for the stop-gradient operator and β is a trade-off weight. The first term of the objective aims to measure the reconstruction loss and the rest terms aim to regularize the high-fidelity prototypes in codebooks. The encoder-decoder network is implemented by an elegant U-Net, detailed in Section 3.4.1. The counselor model is inserted into the first stage of the unfolding framework as the first denoiser and shares the quantized $\mathbf{F}_q$ and framework with denoisers in the later stages.

Phase 2: At the previous phase, the high-fidelity prototypes learned in a self-supervised manner are reckoned as prior knowledge in the successive unfolding stages to reconstruct the high-quality HSI data. Fixing the decoder and codebook, we aim to leverage knowledge for training an unfolding network to reconstruct HSIs from measurement. In the second phase, the codebook and decoder are fixed, while the encoder adapts its parameters to encode $\boldsymbol{x}^l$, which can be recognized as the noisy HSI. Thus, the training process in this phase is relatively easier compared to directly inputting the measurement. Notably, the training objective in the second phase only contains the reconstruction loss for denoiser updating.

3.4 Mask-aware knowledge-assist Denoising

3.4.1 Overview of the Denoiser: Bridge U-Net

Obtained the mask-aware features and high-fidelity prior knowledge, how to make good use of them remains unsolved. Therefore, we propose a U-Net-shaped framework for denoising with information filtering bridges between the hierarchical symmetric encoders and decoders. Instead of using a direct

skip connection, we implement a bridge module including several convolution layers to aggregate information from each spatial level and filter unimportant information. Furthermore, the features from the encoder and the filtered features of a higher-level bridge are concatenated and fused as a 'clean' residual information for decoding. As shown in Figure 3 (a), the multi-level mask features and quantized features are integrated for better reconstruction, implemented with the proposed mask-aware knowledge-assist attention module in transformer encoder and decoders.

3.4.2 LEVERAGING THE MASK AND HIGH-FIDELITY INFORMATION INTO TRANSFORMER

In the Transformer block, we introduce the multi-scale mask features F_M^i and quantized high-fidelity HSI feature $\mathbf{F}_q$ into the vanilla multi-head attention, named Mask-aware Knowledge-assist Attention module (MKA). As shown in the Fig. 3 (b), the MKA encodes input $\mathbf{U}_i$ by 1×1 convolutions and deformable convolutions into query ($\mathbf{Q}$), key ($\mathbf{K}$) and value ($\mathbf{V}$) (with the same size as $\mathbf{U}_i$). Here, we implement a cross attention, *i.e.*modulate $\mathbf{K}$ with selected high-fidelity HSI feature $Sigmoid(Conv(\mathbf{F}_q))$ and modulate $\mathbf{V}$ with mask feature F_M^i while keep $\mathbf{Q}$ unchanged. The Sigmoid and convolution transform $\mathbf{F}_q$ into a proper latent space to interact with $\mathbf{K}$. The MKA can be formulated as:

$$\mathbf{MKA}(\mathbf{U}_i) = W_{c1}\mathbf{V} \odot Softmax(\mathbf{K} \odot \mathbf{Q}/\alpha) \tag{11}$$

$$\mathbf{Q} = \mathbf{W}_{c2}^{Q}(\mathbf{W}_{d}^{Q}Norm(\mathbf{U}_i) + \boldsymbol{b}_{d}^{Q}), \tag{12}$$

$$\mathbf{K} = \mathbf{W}_{c2}^{K}(\mathbf{W}_{d}^{K}Norm(\mathbf{U}_i) + \boldsymbol{b}_{d}^{K}) \odot Sigmoid(\mathbf{W}_{d}^{F_q}\mathbf{F}_q), \tag{13}$$

$$\mathbf{V} = \mathbf{W}_{c2}^{V}(\mathbf{W}_{d}^{V}Norm(\mathbf{U}_i) + \boldsymbol{b}_{d}^{V}) \odot \mathbf{F}_M^i, \tag{14}$$

where $\mathbf{W}_{c1}$ and $\mathbf{W}_{c2}$ represent weights of bias-free convolution, $\mathbf{W}_d, \boldsymbol{b}_d$ are weight and bias of the deformable convolution, $F_M^i, i = 1,2,3$ denotes the mask feature of different spatial levels, and F_q is the quantized feature which refers to the learned prior knowledge. Leveraging mask features enables the encoder and decoder to make the most use of degradation information at each spatial level and align the position of the mask-relative noise on x^k. The high-fidelity knowledge learned from the uncompressed images provides guidance for high-quality reconstruction, as it preserves detailed GT features. In summary, the MKA module helps to focus on relevant information instead of falling into the noisy HSI data itself while considering the information of mask and prior HSI knowledge, investigating relationships among these features.

4 EXPERIMENTS

We conduct experiments on both simulation and real HSI datasets. Following the approaches in Meng et al. (2020b;a); Huang et al. (2021); Cai et al. (2022b), we select a set of 28 wavelengths ranging from 450-650nm by employing spectral interpolation techniques applied to the HSI data.

4.1 EXPERIMENTAL SETTINGS

Simulation and Real Datasets: We adopt two widely used HSI datasets, i.e., CAVE Park et al. (2007) and KAIST Choi et al. (2017) for simulation experiments. The CAVE dataset comprises 32 HSIs with a spatial size of 512 $\times$ 512. The KAIST dataset includes 30 HSIs with a spatial size of 2704 $\times$ 3376. Following previous works Meng et al. (2020b;a); Huang et al. (2021); Cai et al. (2022b), we employ the CAVE dataset as the training set, while 10 scenes from the KAIST dataset are utilized for testing. During the training process, a real mask of size 256 $\times$ 256 pixles is applied. In our real experiment, we utilized the HSI dataset captured by the SD-CASSI system in Meng et al. (2020b). The system captures real-world scenes of size 660 $\times$ 714$\times$ 28 with wavelengths spanning from 450 to 650 nm and dispersion of 54 pixels.

Implementation Details: For the codebook settings, the item number N of the codebook is set to 4096, and the code dimension d is set to 112. In the first phase of training, the trade-off weight $\beta = 0.25$. For all phases of training, we use the Adam Kingma & Ba (2014) optimizer with a batch size of 4. and set the learning rate to 4×10^{-4}. PSNR and SSIM Wang et al. (2004) are utilized as our metrics. Our method is implemented with the PyTorch framework and trained using four NVIDIA RTX3090 GPUs.

Table 1: The average results of PSNR in dB (top entry in each cell), SSIM (bottom entry in each cell) on the 10 synthetic spectral scenes.

Algorithms	Scene1	Scene2	Scene3	Scene4	Scene5	Scene6	Scene7	Scene8	Scene9	Scene10	Avg
TwIST	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	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	23.10 0.584	24.36 0.669
DeSCI	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
HSSP	31.48 0.858	31.09 0.842	28.96 0.823	34.56 0.902	28.53 0.808	30.83 0.877	28.71 0.824	30.09 0.881	30.43 0.868	28.78 0.842	30.35 0.852
DNU	31.72 0.863	31.13 0.846	29.99 0.845	35.34 0.908	29.03 0.833	30.87 0.887	28.99 0.839	30.13 0.885	31.03 0.876	29.14 0.849	30.74 0.863
DGSMP	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
HDNet	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++	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	35.99 0.951
CST-L+	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-9stg	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
DADF-Net (Plus-3)	37.46 0.965	39.86 0.976	41.03 0.974	45.98 0.989	35.53 0.972	37.02 0.975	36.76 0.958	34.78 0.971	40.07 0.976	34.39 0.962	38.29 0.972
RDLUF-MixS2-3stg	36.67 0.953	38.48 0.965	40.63 0.971	46.04 0.986	34.63 0.963	36.18 0.966	35.85 0.951	34.37 0.963	38.98 0.966	33.73 0.950	37.56 0.963
Ours-3stg	36.86 0.959	38.46 0.967	40.83 0.974	45.92 0.988	35.01 0.967	36.43 0.971	36.41 0.958	34.90 0.968	39.31 0.971	33.80 0.956	37.90 0.969
RDLUF-MixS2-9stg	37.94 0.966	40.95 0.977	43.25 0.979	47.83 0.990	37.11 0.976	37.47 0.975	38.58 0.969	35.50 0.970	41.83 0.978	35.23 0.962	39.57 0.974
Ours-9stg	37.98 0.969	41.20 0.981	43.34 0.979	47.70 0.992	37.16 0.977	37.82 0.978	38.61 0.970	36.44 0.978	42.64 0.983	35.29 0.967	39.82 0.977

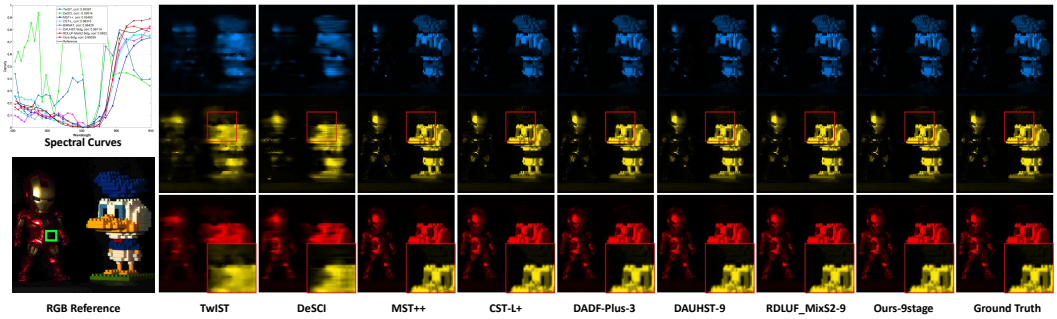

Figure 4: The synthetic data comparisons. 4 out of 28 wavelengths are selected to compare visually. 'Corr' is the correlation coefficient between one method curve and the ground truth curve of the chosen (green box) region. Our method has a more accurate wavelength curve than others.

4.2 Compare with State-of-the-art

We compare our proposed method with recent state-of-the-art (SOTA) methods including deep unfolding series RDLUF-MixS2 Dong et al. (2023), DAUHST Cai et al. (2022c), end-to-end designed networks DADF-Net Xu et al. (2023), CST Cai et al. (2022a), MST Cai et al. (2022b), HDNetHu et al. (2022) and traditional model-based methods TwIST Bioucas-Dias & Figueiredo (2007) and DeSCI Liu et al. (2019) on both synthetic and real datasets. Other methods like GAP-TV Yuan (2016), HSSP Wang et al. (2019), DNU Wang et al. (2020) and DGSMP Huang et al. (2021) are compared on synthetic data.

Synthetic data: Table 1 shows quantitative comparisons on synthetic data that our proposed method outperforms other compared methods. It surpasses the recent deep unfolding method RDLUF-MixS2 both in the 3-stage (+0.34dB) and 9-stage (+0.25 dB) network according to average PSNR and SSIM. The Fig. 4 shows the visual reconstruction results. 4 out of 28 wavelengths are selected

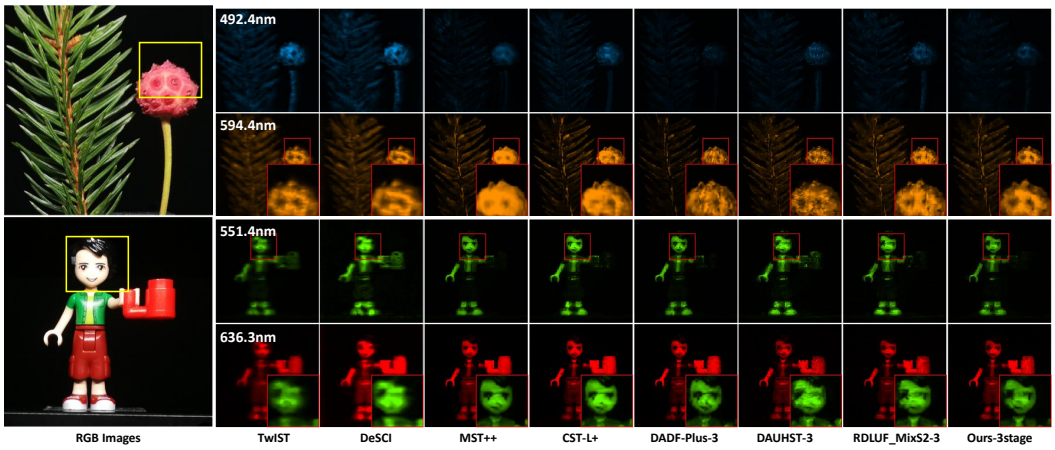

Figure 5: The real data comparisons of 2 scenes. 4 out of 28 wavelengths are selected to compare visually.

to compare. The region within the red box was chosen to analyze the wavelength accuracy. 'Corr' is the correlation coefficient between one method and the ground truth of the chosen region. According to the 'Corr', our method (0.99) has a more accurate wavelength curve than others (0.98 for RDLUF-MixS2). Moreover, as shown in Fig. 1, our method demonstrates a stronger capability to reconstruct intricate details and textures, while previous methods tend to over-smooth the surface. The visualization of the learned knowledge in (b) further illustrates the learned patterns facilitate improved image fidelity. Each patch in (b) denotes a codebook value which is decoded to the natural image domain to represent some basic texture elements.

Real data: Fig. 5 shows reconstruction results of multiple methods on the real scene 'Lego Man' and 'Real Plant'. 4 out of 28 wavelengths are selected to compare visually. Unlike the previous SOTA RDLUF-MixS2, we can see that our method can reconstruct fewer artifacts horizontally and vertically. More results can be seen in the supplementary material (SM).

5 Ablation Study

For the ablation study, we train our model on the synthetic training data with 3 unfolding stage models. Table 5 shows our contributions yield quality performance improvements. We only show main module comparisons here, further ablation study can also be seen in the SM. Row 4 in Table 5 means that it only uses normal denoiser in the first stage unfolding and only trains the network once from measure to reconstruction. $\mathbf{F}_q$ is also removed in MKA. It demonstrates that our two-phase learning strategy leads to a 0.1 dB improvement. Row 1 in Table 5 removes mask encoder and $\mathbf{F}_M^i$ in attention. It demonstrates that our mask cross encoder with cross attention improves the performance by 0.29 dB. Row 2 uses two-phase training but $\mathbf{F}_q$ in attention. It demonstrates that our prior knowledge of cross attention improves the performance by 0.16 dB. Row 3 in Table 5 uses a plain ADMM projection in our unfolding network, dropping the performance by 0.45 dB. It illustrates that our DE-ADMM plays a role in the unfolding network.

Method	PSNR (dB)	Params (M)
w/o MCE and MF	37.61	2.317
w/o HFK in MKA	37.74	2.322
plain ADMM projection	37.45	2.187
w/o two-phase strategy	37.80	2.430
Our Full Model	37.90	2.325

Figure 6: Ablation study of our method.

6 Conclusion

This paper introduces a spectral SCI reconstruction network that leverages high-fidelity knowledge from clean HSIs as well as the cascade degradation information. It achieves state-of-the-art performance on simulated data and real data, surpassing the previous best reconstruction algorithms both in spatial and spectral dimensions. In addition, we proposed a two-phase training strategy within a deep unfolding framework to transfer essential knowledge. We hope to give a new way of modeling high-fidelity knowledge from ground-truth observation space for video compressive, which can potentially extend to a wider range of application scenarios in the future, such as video SCI.

REFERENCES

M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Img. Sci.*, 2(1):183–202, March 2009.

J.M. Bioucas-Dias and M.A.T. Figueiredo. A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Transactions on Image Processing*, 16(12):2992–3004, December 2007.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, January 2011.

Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Coarse-to-fine sparse transformer for hyperspectral image reconstruction. In *European Conference on Computer Vision*, pp. 686–704. Springer, 2022a.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17502–17511, 2022b.

Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc V Gool. Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging. *Advances in Neural Information Processing Systems*, 35:37749–37761, 2022c.

Xun Cao, Tao Yue, Xing Lin, Stephen Lin, Xin Yuan, Qionghai Dai, Lawrence Carin, and David J Brady. Computational snapshot multispectral cameras: Toward dynamic capture of the spectral world. *IEEE Signal Processing Magazine*, 33(5):95–108, 2016.

Stanley H. Chan, Xiran Wang, and Omar A. Elgendy. Plug-and-play ADMM for image restoration: Fixed-point convergence and applications. *IEEE Transactions on Computational Imaging*, 3: 84–98, 2017.

Adam S Charles, Bruno A Olshausen, and Christopher J Rozell. Learning sparse codes for hyperspectral imagery. *IEEE Journal of Selected Topics in Signal Processing*, 5(5):963–978, 2011.

Ziheng Cheng, Bo Chen, Ruiying Lu, Zhengjue Wang, Hao Zhang, Ziyi Meng, and Xin Yuan. Recurrent neural networks for snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2264–2281, 2022.

Inchang Choi, Daniel S. Jeon, Giljoo Nam, Diego Gutierrez, and Min H. Kim. High-quality hyperspectral reconstruction using a spectral prior. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia 2017)*, 36(6):218:1–13, 2017. doi: 10.1145/3130800.3130810. URL `http://dx.doi.org/10.1145/3130800.3130810`.

Yubo Dong, Dahua Gao, Tian Qiu, Yuyan Li, Minxi Yang, and Guangming Shi. Residual degradation learning unfolding framework with mixing priors across spectral and spatial for compressive spectral imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22262–22271, 2023.

Hao Du, Xin Tong, Xun Cao, and Stephen Lin. A prism-based system for multispectral video acquisition. In *2009 IEEE 12th International Conference on Computer Vision*, pp. 175–182. IEEE, 2009.

Alexander FH Goetz, Gregg Vane, Jerry E Solomon, and Barrett N Rock. Imaging spectrometry for earth remote sensing. *science*, 228(4704):1147–1153, 1985.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17542–17551, 2022.

Tao Huang, Weisheng Dong, Xin Yuan, Jinjian Wu, and Guangming Shi. Deep gaussian scale mixture prior for spectral compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16216–16225, 2021.

Shirin Jalali and Xin Yuan. Snapshot compressed sensing: Performance bounds and algorithms. *IEEE Transactions on Information Theory*, 65(12):8005–8024, 2019.

Younghyun Jo and Seon Joo Kim. Practical single-image super-resolution using look-up table. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 691–700, 2021.

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

David Kittle, Kerkil Choi, Ashwin Wagadarikar, and David J Brady. Multiframe image estimation for coded aperture snapshot spectral imagers. *Applied Optics*, 49(36):6824–6833, 2010.

Lu Li, Wei Li, Ying Qu, Chunhui Zhao, Ran Tao, and Qian Du. Prior-based tensor approximation for anomaly detection in hyperspectral imagery. *IEEE Transactions on Neural Networks and Learning Systems*, 33(3):1037–1050, 2020.

Shutao Li, Weiwei Song, Leyuan Fang, Yushi Chen, Pedram Ghamisi, and Jon Atli Benediktsson. Deep learning for hyperspectral image classification: An overview. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9):6690–6709, 2019.

Yang Liu, Xin Yuan, Jinli Suo, David Brady, and Qionghai Dai. Rank minimization for snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12): 2990–3006, Dec 2019.

Patrick Llull, Xuejun Liao, Xin Yuan, Jianbo Yang, David Kittle, Lawrence Carin, Guillermo Sapiro, and David J Brady. Coded aperture compressive temporal imaging. *Optics express*, 21(9):10526–10545, 2013.

Guolan Lu and Baowei Fei. Medical hyperspectral imaging: a review. *Journal of biomedical optics*, 19(1):010901–010901, 2014.

Ruiying Lu, Bo Chen, Ziheng Cheng, and Penghui Wang. Rafnet: Recurrent attention fusion network of hyperspectral and multispectral images. *Signal Processing*, 177:107737, 2020.

Ting Luo, Lishun Wang, and Xin Yuan. Grating-based coded aperture compressive spectral imaging to reconstruct over 190 spectral bands from a snapshot measurement. *Journal of Physics D: Applied Physics*, 56(25):254004, 2023.

Jiawei Ma, Xiao-Yang Liu, Zheng Shou, and Xin Yuan. Deep tensor admm-net for snapshot compressive imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10223–10232, 2019.

Emmanuel Maggiori, Guillaume Charpiat, Yuliya Tarabalka, and Pierre Alliez. Recurrent neural networks to correct satellite image classification maps. *IEEE Transactions on Geoscience and Remote Sensing*, 55(9):4962–4971, 2017.

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

Ziyi Meng, Jiawei Ma, and Xin Yuan. End-to-end low cost compressive spectral imaging with spatial-spectral self-attention. In *European Conference on Computer Vision (ECCV)*, August 2020b.

Ziyi Meng, Jiawei Ma, and Xin Yuan. End-to-end low cost compressive spectral imaging with spatial-spectral self-attention. In *European conference on computer vision*, pp. 187–204. Springer, 2020c.

Ziyi Meng, Xin Yuan, and Shirin Jalali. Deep unfolding for snapshot compressive imaging. *International Journal of Computer Vision*, pp. 1–26, 2023.

Xin Miao, Xin Yuan, Yunchen Pu, and Vassilis Athitsos. λ-net: Reconstruct hyperspectral images from a snapshot measurement. In *IEEE/CVF Conference on Computer Vision (ICCV)*, 2019.

Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239, 2014.

Jong-Il Park, Moon-Hyun Lee, Michael D Grossberg, and Shree K Nayar. Multispectral imaging using multiplexed illumination. In *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8. IEEE, 2007.

Weiqiang Rao, Lianru Gao, Ying Qu, Xu Sun, Bing Zhang, and Jocelyn Chanussot. Siamese transformer network for hyperspectral image target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022.

Ernest Ryu, Jialin Liu, Sicheng Wang, Xiaohan Chen, Zhangyang Wang, and Wotao Yin. Plug-and-play methods provably converge with properly trained denoisers. In *International Conference on Machine Learning*, pp. 5546–5557. PMLR, 2019.

Aziz ul Rehman and Shahzad Ahmad Qureshi. A review of the medical hyperspectral imaging systems and unmixing algorithms' in biological tissues. *Photodiagnosis and Photodynamic Therapy*, 33:102165, 2021.

Burak Uzkent, Aneesh Rangnekar, and Matthew Hoffman. Aerial vehicle tracking by adaptive fusion of hyperspectral likelihood maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 39–48, 2017.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Hien Van Nguyen, Amit Banerjee, and Rama Chellappa. Tracking via object reflectance using a hyperspectral video camera. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pp. 44–51. IEEE, 2010.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

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

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

L. Wang, Z. Xiong, G. Shi, F. Wu, and W. Zeng. Adaptive nonlocal sparse representation for dual-camera compressive hyperspectral imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(10):2104–2111, Oct 2017. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2621050.

Lishun Wang, Zongliang Wu, Yong Zhong, and Xin Yuan. Snapshot spectral compressive imaging reconstruction using convolution and contextual transformer. *Photonics Research*, 10(8):1848–1858, 2022.

Lizhi Wang, Chen Sun, Ying Fu, Min H Kim, and Hua Huang. Hyperspectral image reconstruction using a deep spatial-spectral prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8032–8041, 2019.

Lizhi Wang, Chen Sun, Maoqing Zhang, Ying Fu, and Hua Huang. Dnu: Deep non-local unrolling for computational spectral imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1661–1671, 2020.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Ping Xu, Lei Liu, Haifeng Zheng, Xin Yuan, Chen Xu, and Lingyun Xue. Degradation-aware dynamic fourier-based network for spectral compressive imaging. *IEEE Transactions on Multimedia*, 2023.

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

Xin Yuan, Tsung-Han Tsai, Ruoyu Zhu, Patrick Llull, David Brady, and Lawrence Carin. Compressive hyperspectral imaging with side information. *IEEE Journal of Selected Topics in Signal Processing*, 9(6):964–976, September 2015.

Xin Yuan, Yang Liu, Jinli Suo, and Qionghai Dai. Plug-and-play algorithms for large-scale snapshot compressive imaging. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Xin Yuan, Yang Liu, Jinli Suo, Fredo Durand, and Qionghai Dai. Plug-and-play algorithms for video snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021. doi: 10.1109/TPAMI.2021.3099035.

Shipeng Zhang, Lizhi Wang, Ying Fu, Xiaoming Zhong, and Hua Huang. Computational hyperspectral imaging based on dimension-discriminative low-rank tensor recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10183–10192, 2019.

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.

Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35:30599–30611, 2022.