Latent Diffusion Prior Enhanced Deep Unfolding for Snapshot Spectral Compressive Imaging

Zongliang Wu* Ruiying Lu† Ying Fu‡ Xin Yuan§

Abstract

Snapshot compressive spectral imaging reconstruction aims to reconstruct three-dimensional spatial-spectral images from a single-shot two-dimensional compressed measurement. Existing state-of-the-art methods are mostly based on deep unfolding structures but have intrinsic performance bottlenecks: i) the ill-posed problem of dealing with heavily degraded measurement, and ii) the regression loss-based reconstruction models being prone to recover images with few details. In this paper, we introduce a generative model, namely the latent diffusion model (LDM), to generate degradation-free prior to enhance the regression-based deep unfolding method. Furthermore, to overcome the large computational cost challenge in LDM, we propose a lightweight model to generate knowledge priors in deep unfolding denoiser, and integrate these priors to guide the reconstruction process for compensating high-quality spectral signal details. Numeric and visual comparisons on synthetic and real-world datasets illustrate the superiority of our proposed method in both reconstruction quality and computational efficiency. Code will be released.

1. Introduction

In contrast to normal RGB images which only have three spectral bands, hyperspectral images (HSIs) contain multiple spectral bands with more diverse spectral information. The spectral information serves to characterize distinct objects assisting high-level image tasks [27, 29, 42, 50, 51] and the observation of the world like medical imaging [32, 49] and remote sensing [15, 33]. However, the capture of HSIs is a question that has been studied for a long time because we need to collect HSI signals by 2D sensors. The conventional way of spectral imaging is scanning. It scans the scenes along one dimension such as the spectral or spatial dimension, consuming plenty of time. This way of capture limits the imaging objects to static ones. Thus, for many years, scientists have focused on how to collect HSIs in a quick and convenient method. In 2007, based on compressive sensing theory, a single-shot compressive spectral imaging way [14] was created to efficiently collect HSIs, named coded aperture snapshot spectral imaging (CASSI). The later improvement works [38, 53] provide better imaging quality and lower cost. CASSI modulates the HSI signal across various bands and combine all the modulated spectra to produce a 2D compressed measurement. Consequently, the task of reconstructing the 3D HSI signals from the 2D compressive measurements presents a fundamental challenge for the CASSI system.

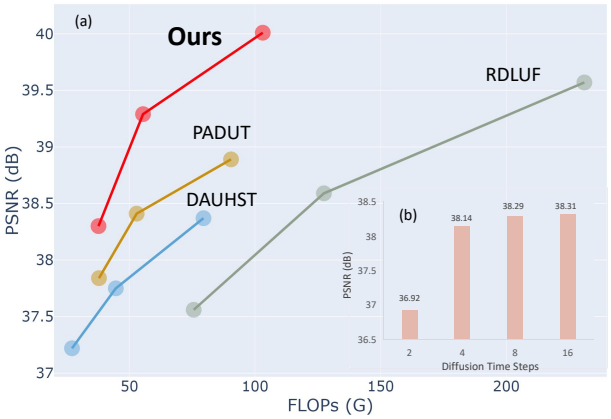

Figure 1. (a) Comparison of PSNR (dB)-FLOPs (G) with previous HSI reconstruction methods. Our proposed method outperforms previous methods using even less computational costs. (b) The ablation study of using different time steps in diffusion. Our method can achieve the desired results requires only very few steps.

The reconstruction process can be viewed as solving an ill-posed problem. Many attempts at solving this problem including traditional model-based methods [1, 2, 60] and the learning-based methods[8, 37, 40] have been proposed since the inception of CASSI system. The deep unfolding network is a combination of convex optimization and neural network prior (denoiser), enjoying both the interpretability of the model-based method and the power of learning-based methods. This branch of methods leads the development trend in recent years [6, 12, 28, 39, 56] and achieves state-of-the-art performance.

However, unlike denoising or reconstruction that recov-

*Zhejiang University, Westlake University
†Xidian University
‡Beijing Institute of Technology
§Westlake University

ers from natural images, CASSI reconstruction has to recover HSIs from the compressed domain measurements, which results in severe degradation according to physical modulation, spectral compression, and unpredictable system noise. Thus, the CASSI reconstruction problem is much harder to learn intrinsic HSI properties than the normal image restoration tasks. In the unfolding framework of the CASSI reconstruction method, the denoising network plays a critical role in deciding the final performance, which is embedded in each stage of the deep unfolding network. However, it always suffers from the performance bottleneck due to the intrinsic ill-posed problem of dealing with heavily degraded measurements. Therefore, a high-performance denoiser with degradation-free knowledge is desired for CASSI reconstruction. Another problem is that previous popular regression-based reconstruction methods have difficulty in recovering details, because the widely used regression losses are conservative with high-frequency details [46].

To address these challenges, we introduce a generative prior in this paper to guide the reconstruction process in an unfolding framework. During training, the prior will be first learned from clean HSIs by an image encoder and then generated by a Latent Diffusion Model (LDM) from Gaussian noise and compressed measurement. Then, the learned prior is embedded into the deep denoiser of the unfolding network by a prior-guided Transformer. Significantly, our unfolding network is able to leverage external prior knowledge from clean HSIs and the powerful generative ability of LDM enhancing its reconstruction performance. The primary contributions presented in this paper can be summarized as follows:

i) We propose a novel **LDM-based unfolding network** for CASSI reconstruction, where the clean image priors are generated by a latent diffusion model to facilitate high-quality hyperspectral reconstruction. To the best of our knowledge, this is the first attempt to combine the physics-driven deep unfolding with generative LDM in CASSI reconstruction.

ii) We design a three-in-one Transformer structure dubbed Trident Transformer (TT) to efficiently extract the correlation among prior knowledge, spatial, and spectral for CASSI reconstruction.

iii) Extensive experiments on the synthetic benchmark and real dataset demonstrate the superior quantitative performances (Fig.1), visual quality (Fig. 2), and lower computational cost of our proposed method.

2. Related Work

2.1. Diffusion Model in Low-level Vision

Diffusion models (DMs) [18, 47] are probabilistic generative models, which model the data distribution by learning a gradual iterative denoising process from the Gaussian distribution to the data distribution. Notably, they demonstrate promising capabilities in generating high-quality samples that encompass a wide range of modes, including super-resolution [13] and inpainting [34]. In light of the impressive achievements of diffusion models in image domains, numerous research endeavors [3, 16, 19, 20] have extended it to video generation. However, diffusion models suffer from significant computation inefficiency regarding data sampling, primarily due to the iterative denoising process required for inference. To address this challenge, several methods propose effective sampling techniques from trained diffusion models [48, 63], or alternatively learning the data distribution from a low-dimensional latent space [44], *i.e*. the latent diffusion model. The latent diffusion has a relatively faster speed and powerful generative ability for super-resolution and inpainting, but similar to the normal diffusion model, it is also prone to issues such as misaligned distribution of fine details and the occurrence of unwanted artifacts, leading to suboptimal performance in distortion-based metrics, e.g., PSNR. Moreover, latent diffusion costs large computational resources both for training and inference due to its large-size encoder and denoiser. Towards this end, some works combine the generative diffusion model with the regression restoration network and work well on distortion-based metrics like deblurring [43]. The recent works [9, 58] employ LDM on many low-level vision tasks and achieve state-of-the-art performance with reasonable computational cost. We name these methods 'integrated diffusion' to distinguish them from the pure diffusion method. However, it is challenging to efficiently employ diffusion models for compressive hyperspectral image reconstruction with heavily degraded measurements.

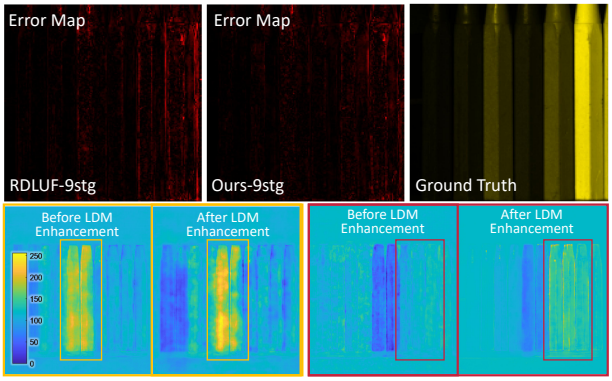

Figure 2. The upper row: the error maps of the previous SOAT and our method. Our method exhibits fewer errors in the edges and textures. The lower row: the feature map (drawn by heatmaps) changes before and after applying LDM enhancement. The enhanced features demonstrate increased concentration on significant parts and edges.

2.2. Hyperspectral Image Reconstruction

Before the advent of the deep learning wave, traditional model-based methods iteratively solved this inverse problem by convex optimization [52, 54, 64] with some handcrafted constraints based on image priors, like sparsity [26] and low-rank [31]. These methods are robust and interpretable but require manual parameter tuning with low reconstruction speed and performance. With the help of deep learning, Plug-and-play (PnP) algorithms [7, 45, 61, 62], embeds pre-trained denoising networks into convex optimization to solve the reconstruction problem, but still has limitations on performance because of the pre-trained denoiser. In recent years, the End-to-end (E2E) reconstruction directly trains a powerful deep neural network, like convolutional neural network (CNN) [10, 22, 33] and Transformers [4, 5], to learn the recovery process from inputs (measurements) to outputs (desired HSIs). However, this simple design lacks interpretability and robustness for various hardware systems. Therefore, an interpretable design of a reconstruction network that unfolds a convex optimization process named deep unfolding is proposed to leverage these problems. A series of CASSI reconstruction works based on deep unfolding [6, 12, 28, 35, 36, 59] are proposed and become the state-of-the-art method. Deep unfolding is able to combine both interpretability in model-based methods and performance in deep learning-based methods to reconstruct CASSI at a fast speed. It changes iterative steps in optimization into several stages in a single network. The prior for optimization becomes a deep neural network denoiser. Since the unfolding needs to define the forward model of imaging, it is also considered as a physics-driven network. However, the recent unfolding networks still have bottlenecks for their regression-based denoiser design and the difficulty of dealing with compressive measurement features. Bearing these concerns, in this work, we propose an 'integrated diffusion' module and integrate it into the physics-driven deep unfolding framework and design an efficient way to aggregate complex features during reconstruction.

3. Proposed Model

3.1. Problem Formulation

The CASSI system has high efficiency in capturing 3D spectral signals by initially coding spectral data with different wavelengths in an aperture and then integrating them into a 2D monochromatic sensor. The mathematical forward process of the widely used single-disperser CASSI (SD-CASSI) [52] can be illustrated as Fig. 3 (a). As can be seen, the original HSI data, denoted as $\boldsymbol{X} \in \mathbb{R}^{W\times H\times N_\lambda}$, is coded by the physical mask $\boldsymbol{M} \in \mathbb{R}^{W\times H}$, where W, H, and N_λ denote the width, height, and the number of spectral channels, respectively. The coded HSI data cube is represented as $\mathbf{X}'(:,:,n_\lambda) = \mathbf{X}(:,:,n_\lambda) \odot \mathbf{M}, n_\lambda = 1,2,\ldots,N_\lambda$, 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 $\mathbf{X}'' \in \mathbb{R}^{W\times \tilde{H}\times N_\lambda}$, where $\tilde{H} = H + d_\lambda$. Here, d_λ represents the shifted distance of the N_λ-th wavelength. This process can be formulated as modulating the shifted version $\tilde{\mathbf{X}} \in \mathbb{R}^{W\times \tilde{H}\times N_\lambda}$ with a shifted mask $\tilde{\mathbf{M}} \in \mathbb{R}^{W\times \tilde{H}\times N_\lambda}$, where $\tilde{\boldsymbol{M}}(i,j,n_\lambda) = \boldsymbol{M}(w,h+d_\lambda)$. At last, the imaging sensor captures the shifted image into a 2D measurement $\mathbf{Y}$, calculated as

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

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

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

where $\boldsymbol{A} \in \mathbb{R}^{W\tilde{H}\times W\tilde{H}N_\lambda}$ denotes the sensing matrix (coded aperture) which is a concatenation of diagonal matrices, that is $\boldsymbol{A} = [\mathbf{D}_1,\ldots,\mathbf{D}_\lambda]$, where $\mathbf{D}_\lambda = Diag(vec(\tilde{\mathbf{M}}(:,:,n_\lambda)))$ is the diagonal matrix with $vec(\tilde{\mathbf{M}}(:,:,n_\lambda))$ as the diagonal elements. In this paper, we will propose a method to solve the ill-posed problem, reconstructing the HSI $\boldsymbol{x}$ from the compressed measurement $\boldsymbol{y}$.

3.2. The Unfolding GAP Framework

Eq. (2) can be typically solved by convex optimization through the following objective:

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

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

To solve the optimization problem, we employ GAP (Generalized Alternating Projection) as our optimization framework, which extends classical alternating projection to the case in which projections are performed between convex sets that undergo a systematic sequence of changes. It can be interrupted anytime to return a valid solution and resumed subsequently to improve the solution [30]. This property is very suitable for deep unfolding which has very limited 'optimization iterations' (stages in the deep unfolding). Specifically, we introduce an auxiliary parameter $\boldsymbol{v}$, Eq. (3) can be written as:

$$(\hat{\boldsymbol{x}},\hat{\boldsymbol{v}}) = \arg\min_{\boldsymbol{x},\boldsymbol{v}} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{v}\|_2^2 + \lambda R(\boldsymbol{v}), \quad \text{s.t.} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}. \tag{4}$$

Then, the problem can be solved by the following subproblems: Firstly, we aim at updating $\boldsymbol{x}$:

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{v}^{(k)} + \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)^{-1}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{v}^{(k)}). \tag{5}$$

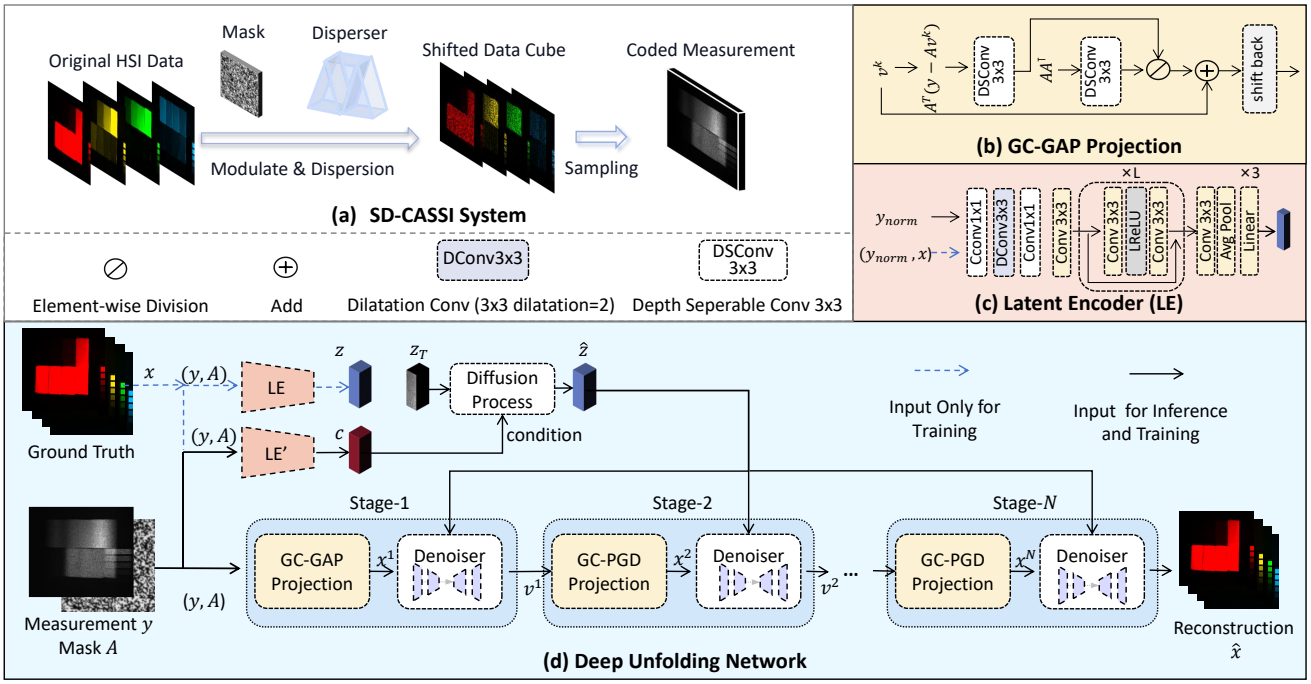

Figure 3. (a) The single disperser CASSI imaging process. HSI data cube is captured by a monochromatic sensor. (b) GC-GAP projection. (c) Latent encoder. (d) Our deep unfolding framework of N stages with the priors generated by diffusion model.

This step projects measurement to a 3D signal space by Euclidean projection. Secondly, we aim at updating v:

$$\boldsymbol{v}^{(k+1)} = \mathscr{D}_{k+1}(\boldsymbol{x}^{(k+1)}), \tag{6}$$

where $\mathscr{D}_k$ is the neural network denoiser of the $k-th$ stage, and λ is the noise penalty factor considered in learned denoisers. This step tries to map $\boldsymbol{x}^{(k+1)}$ to the target signal domain. Considering the projection step Eq. (5), assisted by deep network, we can modify it as follows:

$$\boldsymbol{x}^{k+1} = \boldsymbol{v}^{(k)} + \mathrm{DSC}(\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)^{-1}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{v}^{(k)})), \tag{7}$$

where DSC$(\cdot)$ denotes a set of depthwise separable convolution and GELU [17] operations. The overall unfolding framework is shown in Fig. 3(d), where mask A and measurement y are inputs of the network. According to the Eqs. (5) and (6), the first stage outputs $\boldsymbol{v}^1$ can be obtained. The detailed process is shown in Fig. 3(b). Typically, this process could make use of the position-specific degradation information in the projection part and close the gap between the sensing matrix and degradation matrix [12, 28]. Considering the stage number is much less than the iteration numbers in traditional model-based methods, it is difficult to achieve convergence with limited steps of gradient descent. However, these learnable linear and non-linear variations can be used to correct the gradients in these limited stage unfolding projections, and thus we name it Gradient Correlation GAP (GC-GAP).

3.3. Latent Diffusion Prior Assisted Unfolding Denoising

The denoising process in deep unfolding leads to a natural performance bottleneck due to the intrinsic problem of heavily degraded input. Thus, we introduce external degradation-free prior knowledge to adaptively compensate for the denoising process. We will then introduce this process in a two-phase manner.

Phase I: Learning Prior Knowledge from clean HSIs. In this phase, we use an image encoder to compress both compressive measurement $\boldsymbol{y}$ and clean HSIs (Ground-Truth hyperspectral images) $\boldsymbol{x}$ into latent space. However, instead of simply using measurement $\boldsymbol{y}$, we transfer $\boldsymbol{y}$ by Euclidean projection to 3D HSIs space and normalize it by sensing matrix $\boldsymbol{y}_{norm} \in \mathbb{R}^{W\times H\times N_\lambda} = \boldsymbol{A}^\top\left(\boldsymbol{A}\boldsymbol{A}^\top\right)^{-1}\boldsymbol{y}$. This will improve the balance between two different inputs and easier for the encoder to learn their relation. The input of the encoder in the first phase is $\boldsymbol{I} \in \mathbb{R}^{W\times H\times 2B} =$ concatenate$(\boldsymbol{y}_{norm}, \boldsymbol{x})$. Thus the latent encoder process can be written as $\boldsymbol{z}_{GT} \in \mathbb{R}^{N\times C} = \mathrm{LE}(\boldsymbol{I})$, where $N \ll W \times H$, C is the latent feature channel number. The LE can be seen in Fig. 3 (c), it consists of several residual convolution and linear operations. Then this learned representation $\boldsymbol{z}_{GT}$ will be used as prior in the denoiser to compensate for the denoising errors. The Deep Unfolding Network (DUN) will reconstruct HSI signals using measurement and mask with the assistance of $\boldsymbol{z}_{GT}$, *i.e.* $\hat{\boldsymbol{x}} = \mathrm{DUN}(\boldsymbol{y}, \boldsymbol{A}, \boldsymbol{z}_{GT})$. In this phase, we only use the reconstruction loss: $\mathscr{L}_{\text{rec}} = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_1$.

Phase II: Generating Prior by Latent Diffusion Model.

After learning the prior representation from clean HSIs, we aim to learn an LDM to generate this prior condition on measurement $\boldsymbol{y}$ in the second phase. Specifically, the encoder in the first phase LE is fixed to encode clean HSIs and measurements to $\boldsymbol{z}$ as the generative object of latent space, *i.e*. the starting point of the forward Markov process in the diffusion model. Then as usual forward process, Gaussian noise will be gradually added on $\boldsymbol{z}$ across T time steps according to the parameter β_t, stated as:

$$\begin{aligned} q\left(\boldsymbol{z}_{1:T} \mid \boldsymbol{z}_0\right) &= \textstyle\prod_{t=1}^{T} q\left(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}\right), \forall t=1,\ldots,T, \\ q\left(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}\right) &= \mathcal{N}\left(\boldsymbol{z}_t; \sqrt{1-\beta_t}\boldsymbol{z}_{t-1}, \beta_t \mathbf{I}\right), \end{aligned} \tag{8}$$

where $\boldsymbol{z}_t$ represents the noisy features at the t-th step, and $\boldsymbol{z}_0 = \boldsymbol{z}_{GT}$ is the generative target. $\beta_{1:T} \in (0,1)$ are hyperparameters that control the variance of the Gaussian distribution $\mathcal{N}$. Through iterative derivation with reparameterization [25], Eq. equation 8 can be written as:

$$\begin{aligned} q\left(\boldsymbol{z}_t \mid \boldsymbol{z}_0\right) &= \mathcal{N}\left(\boldsymbol{z}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0, \left(1-\bar{\alpha}_t\right)\mathbf{I}\right), \\ \alpha &= 1-\beta_t, \quad \bar{\alpha}_t = \textstyle\prod_{i=1}^{t} \alpha_i. \end{aligned} \tag{9}$$

The reverse process involves generating the prior features from a pure Gaussian distribution step-by-step condition on the measurement. The reverse process operates as a T-step Markov chain that runs backward from $\boldsymbol{z}_T$ to $\boldsymbol{z}_0$. Specifically, the posterior distribution of the reverse step from $\boldsymbol{z}_t$ to $\boldsymbol{z}_{t-1}$ can be formulated as:

$$\begin{aligned} q\left(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t, \boldsymbol{z}_0\right) &= \mathcal{N}\left(\boldsymbol{z}_{t-1}; \boldsymbol{\mu}_t\left(\boldsymbol{z}_t, \boldsymbol{z}_0\right), \tfrac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \mathbf{I}\right), \\ \boldsymbol{\mu}_t\left(\boldsymbol{z}_t, \boldsymbol{z}_0\right) &= \tfrac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{z}_t - \tfrac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right), \end{aligned} \tag{10}$$

where $\boldsymbol{\epsilon}$ represents the noise added on $\boldsymbol{z}_t$, which is the only uncertain variable. Thus, we adopt a denoising network, denoted as $\boldsymbol{\epsilon}_\theta$, to estimate the noise $\boldsymbol{\epsilon}$ at each step, following the previous works [18, 44, 46]. Considering our diffusion model operates in the latent space, we utilize another latent encoder to extract latent features, denoted as LE$'$, with the same structure as the latent encoder of phase I. Specifically, LE$'$ compresses the normalized measurement $\boldsymbol{y}_{norm}$ into latent space to get the condition latent $\mathbf{c} \in \mathbb{R}^{N \times C'}$. In the end, we use the denoising network to predict the noise ε_t according to $\boldsymbol{z}_t$ of the previous step in reverse process and the condition features $\mathbf{c}$, stated as $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}_\theta\left(\boldsymbol{z}_t, \mathbf{c}, t\right)$. With the substitution of $\boldsymbol{\epsilon}_\theta$ in Eq. (10) and set the variance as $1-\alpha_t$, the reverse inference can be stated as:

$$\boldsymbol{z}_{t-1} = \tfrac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{z}_t - \tfrac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta\left(\boldsymbol{z}_t, \mathbf{c}, t\right)\right) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t, \tag{11}$$

where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \mathbf{I})$. Finally, we can generate the target prior feature $\hat{\boldsymbol{z}} \in \mathbb{R}^{N \times C'}$ after T iterative sampling $\boldsymbol{z}_t$ by Eq. (11). As shown in Fig. 3(d), the predicted prior feature is then used to guide the Transformer in denoiser. Notably, since the distribution of the latent space with the size of $\mathbb{R}^{N \times C}$ (e.g, 16×256) is much simpler than that of images with size $\mathbb{R}^{H \times W \times B}$ (e.g, $256 \times 256 \times 28$), the prior feature can be generated with a small number of iterations T, corresponding to paper [44].

Typically, training the diffusion model refers to training the denoising network $\boldsymbol{\epsilon}_\theta$. Following the previous works [18, 48], we train the model by optimizing the weighted variational bound. The training objective is:

$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\boldsymbol{z} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \mathbf{c}, t\right)\right\|_2^2, \tag{12}$$

where $\boldsymbol{z}$ and $\mathbf{c}$ are ground-truth prior features and the latent condition representations defined above; $t \in [1, T]$ is the randomly sampled time step; $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ denotes the sampled Gaussian noise. We jointly update all the parameters in the network with the objective loss function of the second phase, including: the deep unfolding network, the feature encoder LE$'$, and the diffusion denoising network $\boldsymbol{\epsilon}_\theta$. The objective loss function of the second phase can be stated as:

$$\mathscr{L}_{\text{diff}} = \|\hat{\boldsymbol{z}} - \boldsymbol{z}\|_1, \quad \mathscr{L}_{\text{all}} = \mathscr{L}_{\text{rec}} + \mathscr{L}_{\text{diff}}. \tag{13}$$

3.4. Aggregate Features by Trident Transformer

Previous HSI reconstruction methods usually only exploit the relation between spatial and spectral, both externally and internally. However, the spatial-spectral relations are challenging to explore only with compressed measurements. Therefore, we design a Transformer, named Trident Transformer (TT), to effectively aggregate our learned high-quality degradation-free prior knowledge for compensation.

Firstly, inspired by the multi-scale operations in previous papers [9, 58] with hierarchical structures, we downsample the prior to obtain the multi-scale prior representations. Specifically, three downsampling layers are employed, and the outputs contain prior features of three scales, stated as:

$$\boldsymbol{z}^i = \begin{cases} \boldsymbol{z}_{GT}, & if \quad i=1, \\ \text{downsample}(\boldsymbol{z}^{i-1}), & if \quad i>1 \end{cases}, \tag{14}$$

where $\boldsymbol{z}^i \in \mathbb{R}^{\frac{N}{2^{i-1}} \times 2^{i-1}C}$, $i = 1,2,3$. For $i=1$, $\boldsymbol{z}^i = \boldsymbol{z}_{GT}$, which is computed in the first phase training; For $i = 2,3$, $\boldsymbol{z}^i = \hat{\boldsymbol{z}}$, which is utilized for training and inference in the second phase.

As shown in Fig. 4, our Trident Transformer includes three branches: spatial flow, cross-spectral flow, and cross-prior flow. Each branch shares the information flow with others and is then fused by the aggregation layer and a feed-forward network (FFN). Before the embedding layer, the input feature at i-th scale $\mathbf{U}_i \in \mathbb{R}^{H_i \times W_i \times C_i}$ is split into $\mathbf{U}_i^C \in \mathbb{R}^{H_i \times W_i \times \frac{C_i}{2}}$ and $\mathbf{U}_i^S \in \mathbb{R}^{H_i \times W_i \times \frac{C_i}{2}}$ along the channel dimension, denoting cross flow input and spatial flow input respectively. The spatial flow consists a series of mobile

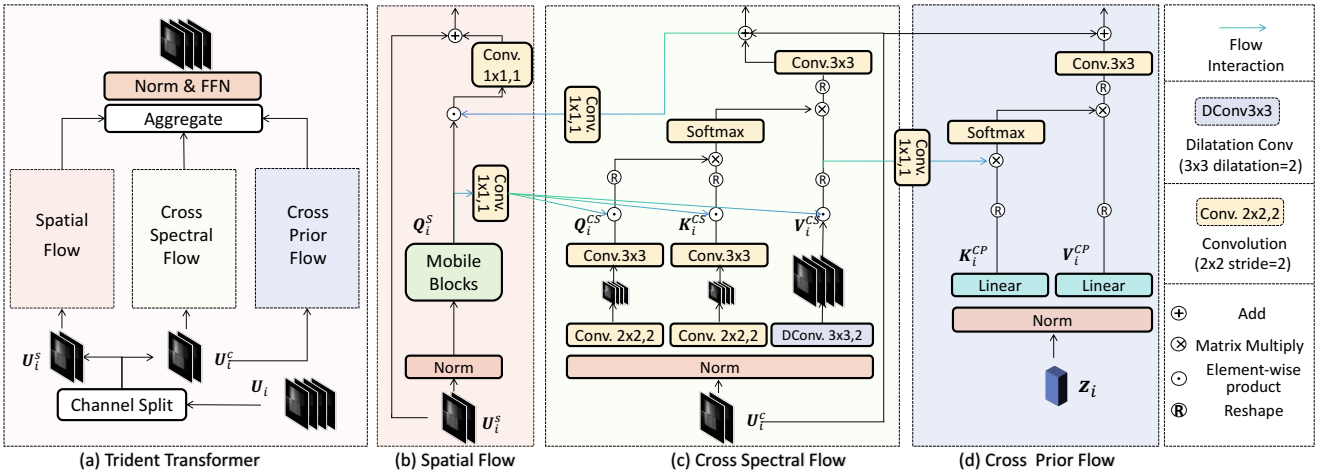

Figure 4. The illustration of (a) the overall structure of Trident Transformer, and (b)-(d) the detailed sub-modules in the Trident Transformer.

blocks [21] without batch norms, which aims to extract representative visual information with fewer parameters and FLOPs.

3.4.1 Cross Spectral Flow

In cross spectral flow (CSF) module, as shown in Fig. 4 (c), we design a shrunken multi-head self-attention with cross-scale embedding. This flow primarily focuses on the spectral dimension and aims to save computational burden acoording to the spatial size. Specifically, we compress the spatial size of the query embedding ($\mathbf{Q}$) and key embedding ($\mathbf{K}$) to $1/4$ and expand its channel twice. After establishing the spectral correlation, we use a dilation convolution on the value embedding ($\mathbf{V}$) to obtain larger perceptual field information with expanded channel dimension and unchanged spatial dimension. The CSF can be formulated as:

$$MSA_i^{CS}(\mathbf{U}_i) = \mathbf{W}_{c1}^{CS}\mathbf{V}_i^{CS} \odot \text{Softmax}(\mathbf{K}_i^{CS} \odot \mathbf{Q}_i^{CS}/\alpha), \quad (15)$$

$$\mathbf{Q}_i^{CS} = \mathbf{W}^{QCS}\mathbf{U}_i, \ \mathbf{K}_i^{CS} = \mathbf{W}^{KCS}\mathbf{U}_i, \ \mathbf{V}_i^{CS} = \mathbf{W}^{VCS}\mathbf{U}_i, \quad (16)$$

where $\mathbf{W}^*$ represent the weights of bias-free convolution.

3.4.2 Cross Prior Flow

Cross prior flow (CPF) in Fig. 4 (d) is a variable shared multi-head cross-attention. The query in this flow is borrowed from the value of CSF which is extracted from a large perceptive field with more spatial information. In this way, the prior could facilitate to compensate for spatial deficiency. Compared to the spectral recovery, the spatial recovery is typically more challenging. Our manipulation can be formulated as:

$$MSA_i^{CP}(\mathbf{U}_i) = W_{c1}\mathbf{V} \odot \text{Softmax}(\mathbf{K} \odot \mathbf{Q}/\alpha), \quad (17)$$

$$\mathbf{Q}_i^{CP} = \mathbf{Q}_i^{CS}, \ \mathbf{K}_i^{CP} = \mathbf{W}_z^K \boldsymbol{z}_i, \ \mathbf{V}_i^{CP} = \mathbf{W}_z^V \boldsymbol{z}_i, \quad (18)$$

where $\boldsymbol{z}^i, i = 1,2,3$ denotes the prior feature of different spatial levels.

3.4.3 Flow Interaction and Aggregation

In order to compensate for the deficiency of spatial information in CSF and CPF, and the spectral information in the spatial flow, we fuse the compensation information together to reconstruct hyperspectra images. As shown in Fig. 4, the colorful arrows represents the information interactions between each flow. Specifically, information of each module modulate with other flows, where the 1×1 convolutions serve as compensation bridges. The aggregation part consists of concatenation, convolution layers, and an activation function for a weighted combination of each flow output. In our Trident Transformer, the prior knowledge learned from the clean images will provide compensation for reconstruction in both spatial and spectral details, avoiding the influence of degraded measurements.

4. Experiments

We conduct experiments on both simulation and real HSI datasets. Following the approaches in Cai et al. [5], Huang et al. [23], Meng et al. [36, 37], 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 [41] and KAIST [11] for simulation experiments. The CAVE dataset comprises 32 HSIs with a spatial size of 512 × 512. The KAIST dataset includes 30 HSIs with a spatial size of 2704 × 3376. Following previous works [5, 23, 36, 37], we employ the CAVE dataset as the training set, while 10 scenes from the KAIST dataset are utilized for testing. During the training process,

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
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
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
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
BIRNAT	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-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
PADUT-5stg	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
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	37.08 0.961	39.53 0.974	41.67 0.978	46.37 0.991	35.73 0.972	36.71 0.973	36.94 0.962	35.13 0.971	39.96 0.976	33.96 0.961	38.31 0.972
Ours-5stg	37.82 0.967	40.45 0.979	43.25 0.982	47.65 0.993	36.99 0.978	37.22 0.977	38.13 0.970	35.79 0.976	41.74 0.982	34.83 0.965	39.38 0.977
PADUT-12stg	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
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	38.07 0.969	41.16 0.982	43.70 0.983	48.01 0.993	37.76 0.980	37.65 0.980	38.58 0.973	36.31 0.979	42.66 0.984	35.18 0.967	39.91 0.979
Ours-10stg	38.18 0.970	41.29 0.983	43.65 0.983	47.87 0.994	37.94 0.981	37.72 0.980	38.94 0.974	36.40 0.979	42.94 0.985	35.16 0.968	40.01 0.980

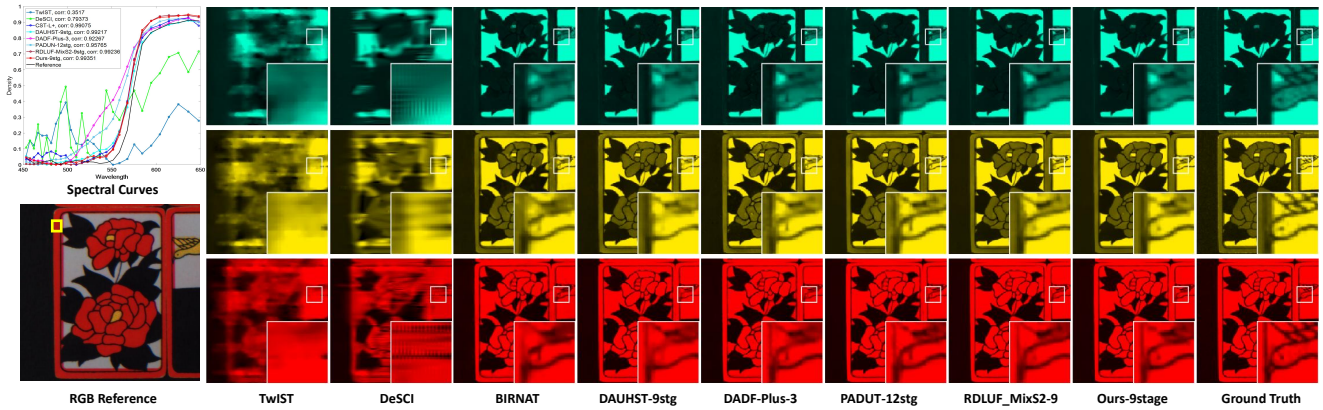

Figure 5. The visualization result on synthetic data. 3 out of 28 wavelengths are selected for visual comparison. 'Corr' in the top left curve is the correlation coefficient between one method curve and the ground truth curve of the chosen (golden box) region.

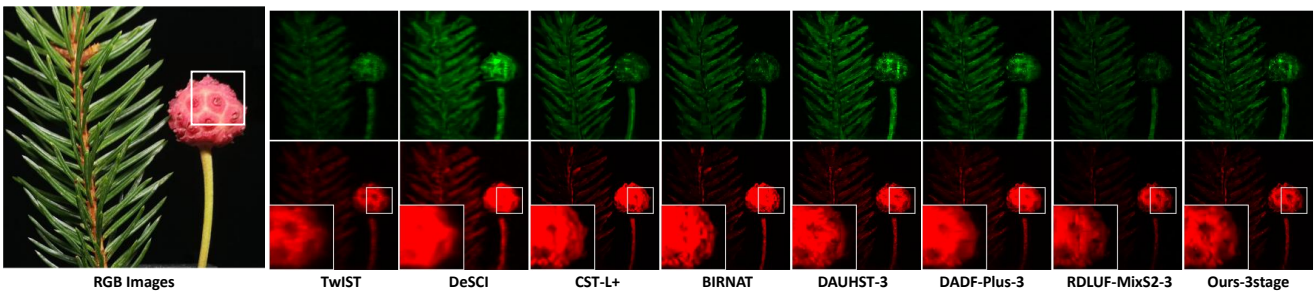

Figure 6. The real data comparisons. 2 out of 28 wavelengths are selected for visual comparison.

Method	PSNR (dB)	FLOPs (G)	Time (ms/10scenes)
Our Full Model	38.31	37.71	421.96
w/o prior	37.63	36.66	403.90
w/o prior and TT	38.11	31.60	305.55

Table 2. Ablation study of our method. 'Time' denotes the total inference time of each method dealing with 10 synthetic test scenes.

a real mask of size 256 $\times$ 256 pixels is applied. In our real experiment, we utilized the HSI dataset captured by the SD-CASSI system in Meng et al. [37]. 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 diffusion settings, the iteration number T of the diffusion is set to 16, and the latent space dimension N is set to 256. For all phases of training, we use the Adam [24] optimizer and set the learning rate to 4×10^{-4}. PSNR and SSIM [57] are utilized as our metrics. Our method is implemented with the PyTorch and trained using NVIDIA RTX3090 GPUs. More details can be seen in the supplementary material (SM).

4.2. Compare with State-of-the-art

We numerically compare our proposed method with previous methods including the end-to-end networks: DADF-Net [59], CST [4], BIRNAT [10]; the deep unfolding methods: RDLUF-MixS2 [12], PADUT [28], DAUHST [6], DNU [55]; and traditional model-based methods: TwIST [2] and DeSCI [31]. The visual comparisons on both synthetic and real datasets are also conducted with recent state-of-the-art (SOTA) methods.

Synthetic data: The numeric comparisons on synthetic data can be seen in Table 1. Our proposed method surpasses the recent SOTA method RDLUF-MixS2 (+0.44 dB) according to average PSNR and SSIM. Fig. 5 shows the visual reconstruction results. Three wavelengths including striking colors in RGB reference red, yellow, and green are selected to compare. The golden box part in the reference was chosen to calculate and compare the wavelength accuracy. The accuracy metric is the correlation coefficient with the ground truth of the chosen region, *i.e.* the 'Corr' in the curves. According to the 'Corr', our method (0.9935) has a more accurate wavelength curve than others. The zoomed part in the figure also demonstrates that our method has edges of bird wings and bird eyes are clearer than others.

Real data: Two scenes of real SD-CASSI measurement reconstruction results are shown in Fig. 6, and two obvious color regions in each scene RGB references are selected to compare. We can see that our method can reconstruct more details such as the light and dark intricacies of the flower.

5. Ablation Study

In this ablation study, we train our model on the synthetic training data with 3 unfolding stage models. The results are summarized in Table 2, where 'w/o prior' denotes that we only use measurement as the latent encoder input instead of clean HSIs, 'w/o prior, and TT' denotes that the latent encoder, diffusion, and prior flow in the Trident Transformer are removed. The training process only conducts a simple one-phase end-to-end strategy instead of a two-phase training. The ablation illustrates that with LDM prior assistance, we can achieve better reconstruction results, and our design of the Trident Transformer successfully aggregates three types of information and effectively compensates for some reconstruction defects. Fig. 2 also visualizes the feature map changes before and after prior enhancement. The enhanced features demonstrate increased concentration on significant parts and edges. Moreover, without accurate guidance, the LDM will even harm the reconstruction. We also compare the influence of diffusion time steps in Fig. 1, 16 steps are enough for good reconstruction results. The inference time in Table 2 illustrates that time is still in a reasonable range even using diffusion 16 steps.

6. Conclusion

In this paper, we introduce a novel spectral reconstruction network that leverages prior knowledge from the latent diffusion model. The network uses a Trident Transformer to effectively combine physics-driven deep unfolding and powerful latent diffusion model. It achieves state-of-the-art performance on both simulated data and real data and has considerable improvements in computational efficiency. This new strategy of using deep unfolding can shed light on other low-level vision and computational imaging tasks.

References

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

[2] 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, 2007. 1, 8

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2

[4] 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*, pages 686–704. Springer, 2022. 3, 8

[5] 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*, pages 17502–17511, 2022. 3, 6

[6] 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, 2022. 1, 3, 8

[7] 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. 3

[8] 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. 1

[9] Zheng Chen, Yulun Zhang, Ding Liu, Bin Xia, Jinjin Gu, Linghe Kong, and Xin Yuan. Hierarchical integration diffusion model for realistic image deblurring. *arXiv preprint arXiv:2305.12966*, 2023. 2, 5

[10] 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. 3, 8

[11] 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. 6

[12] 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*, pages 22262–22271, 2023. 1, 3, 4, 8

[13] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10021–10030, 2023. 2

[14] M. E. Gehm, R. John, D. J. Brady, R. M. Willett, and T. J. Schulz. Single-shot compressive spectral imaging with a dual-disperser architecture. *Optics Express*, 15(21):14013–14027, 2007. 1

[15] Alexander FH Goetz, Gregg Vane, Jerry E Solomon, and Barrett N Rock. Imaging spectrometry for earth remote sensing. *science*, 228(4704):1147–1153, 1985. 1

[16] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *Advances in Neural Information Processing Systems*, 35:27953–27965, 2022. 2

[17] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 5

[19] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[20] Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. *arXiv preprint arXiv:2206.07696*, 2022. 2

[21] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 6

[22] 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*, pages 17542–17551, 2022. 3

[23] 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*, pages 16216–16225, 2021. 6

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 8

[25] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. 5

[26] 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. 3

[27] 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. 1

[28] Miaoyu Li, Ying Fu, Ji Liu, and Yulun Zhang. Pixel adaptive deep unfolding transformer for hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12968, 2023. 1, 3, 4, 8

[29] 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. 1

[30] Xuejun Liao, Hui Li, and Lawrence Carin. Generalized alternating projection for weighted-2,1 minimization with applications to model-based compressive sensing. *SIAM Journal on Imaging Sciences*, 7(2):797–823, 2014. 3

[31] 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, 2019. 3, 8

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

[33] Ruiying Lu, Bo Chen, Ziheng Cheng, and Penghui Wang. Rafnet: Recurrent attention fusion network of hyperspectral and multispectral images. *Signal Processing*, 177:107737, 2020. 1, 3

[34] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 2

[35] 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*, pages 10223–10232, 2019. 3

[36] Ziyi Meng, Shirin Jalali, and Xin Yuan. Gap-net for snapshot compressive imaging. *arXiv preprint arXiv:2012.08364*, 2020. 3, 6

[37] 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)*, 2020. 1, 6, 8

[38] 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*, pages 187–204. Springer, 2020. 1

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

[40] 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. 1

[41] 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*, pages 1–8. IEEE, 2007. 6

[42] 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. 1

[43] Mengwei Ren, Mauricio Delbracio, Hossein Talebi, Guido Gerig, and Peyman Milanfar. Multiscale structure guided diffusion for image deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10721–10733, 2023. 2

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 5

[45] 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*, pages 5546–5557. PMLR, 2019. 3

[46] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022. 2, 5

[47] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[48] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 5

[49] 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. 1

[50] 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*, pages 39–48, 2017. 1

[51] 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*, pages 44–51. IEEE, 2010. 1

[52] 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. 3

[53] 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

[54] 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, 2017. 3

[55] 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*, pages 1661–1671, 2020. 8

[56] 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. 1

[57] 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. 8

[58] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. *arXiv preprint arXiv:2303.09472*, 2023. 2, 5

[59] 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. 3, 8

[60] Xin Yuan. Generalized alternating projection based total variation minimization for compressive sensing. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2539–2543, 2016. 1

[61] 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)*, 2020. 3

[62] 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*, pages 1–1, 2021. 3

[63] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022. 2

[64] 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*, pages 10183–10192, 2019. 3