Dual Prior Unfolding for Snapshot Compressive Imaging

Jiancheng Zhang [1,*], Haijin Zeng [2,*], Jiezhang Cao [3], Yongyong Chen [4,†], Dengxiu Yu [1], Yin-Ping Zhao [1,†]
[1] Northwestern Polytechnical University, [2] IMEC-UGent, [3] ETH Zürich,
[4] Harbin Institute of Technology (Shenzhen)

Abstract

Recently, deep unfolding methods have achieved remarkable success in the realm of Snapshot Compressive Imaging (SCI) reconstruction. However, the existing methods all follow the iterative framework of a single image prior, which limits the efficiency of the unfolding methods and makes it a problem to use other priors simply and effectively. To break out of the box, we derive an effective Dual Prior Unfolding (DPU), which achieves the joint utilization of multiple deep priors and greatly improves iteration efficiency. Our unfolding method is implemented through two parts, i.e., Dual Prior Framework (DPF) and Focused Attention (FA). In brief, in addition to the normal image prior, DPF introduces a residual into the iteration formula and constructs a degraded prior for the residual by considering various degradations to establish the unfolding framework. To improve the effectiveness of the image prior based on self-attention, FA adopts a novel mechanism inspired by PCA denoising to scale and filter attention, which lets the attention focus more on effective features with little computation cost. Besides, an asymmetric backbone is proposed to further improve the efficiency of hierarchical self-attention. Remarkably, our 5-stage DPU achieves state-of-the-art (SOTA) performance with the least FLOPs and parameters compared to previous methods, while our 9-stage DPU significantly outperforms other unfolding methods with less computational requirement. https://github.com/ZhangJC-2k/DPU

1. Introduction

The advent of compressed sensing has introduced a hardware encoder known as Snapshot Compressive Imaging (SCI) [16, 30, 38]. This encoder offers characteristics like low bandwidth, rapid acquisition, and high data throughput, garnering substantial attention in the domains of low-level vision and computational imaging. SCI employs a two-dimensional (2D) detector to capture modulated three-dimensional (3D) hyperspectral images (HSIs) through

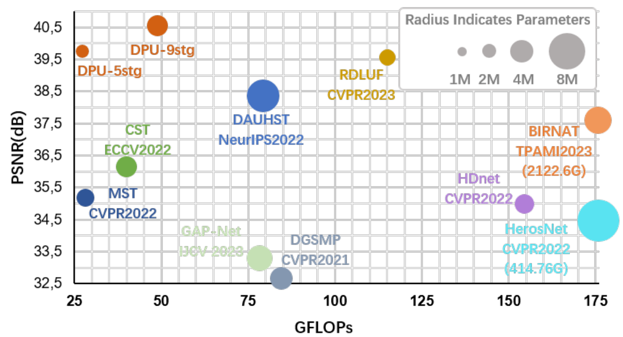

Figure 1. The PSNR-FLOPs-Params analysis comparing the proposed Dual Prior Unfolding (DPU) with latest state-of-the-art methods. Notably, our DPU achieves superior performance while demanding cheaper FLOPs and Parameters, making it a more cost-effective solution for spectral SCI reconstruction.

snapshot measurements [48]. Equipped with this hardware encoder, the development of a high-quality algorithmic decoder becomes imperative for practical SCI systems.

In response to this challenge, both model-based [1, 23, 24, 36, 42, 46] and learning-based [3, 4, 6, 18, 28, 33, 44, 45, 54] approaches have been specifically designed. Notably, deep unfolding methods [5, 14, 19, 32, 41, 52], leveraging deep networks as image priors in iterative algorithms and then implementing end-to-end training, have demonstrated notable success. As researchers pay attention to the important role of other priors such as mediation knowledge [3] and degradation information [5, 14] in the imaging process, the single image prior design can no longer meet their requirements. More and more frameworks such as GAP [32], DAUF [5], and RDLF [14] have been proposed to consider other prior information to assist reconstruction. However, the existing unfolding methods all follow the iterative framework of a single image prior, which limits the efficiency of the unfolding methods and makes it a problem to use other priors simply and effectively. In addition, previous methods [3, 5, 14] identify the critical role of mask degradation while ignoring the effects of shift and compression degradation in the imaging process.

Considering the need for more effective utilization of image priors and degradation-associated priors, we have de-

*Equal Contribution. † Corresponding Authors

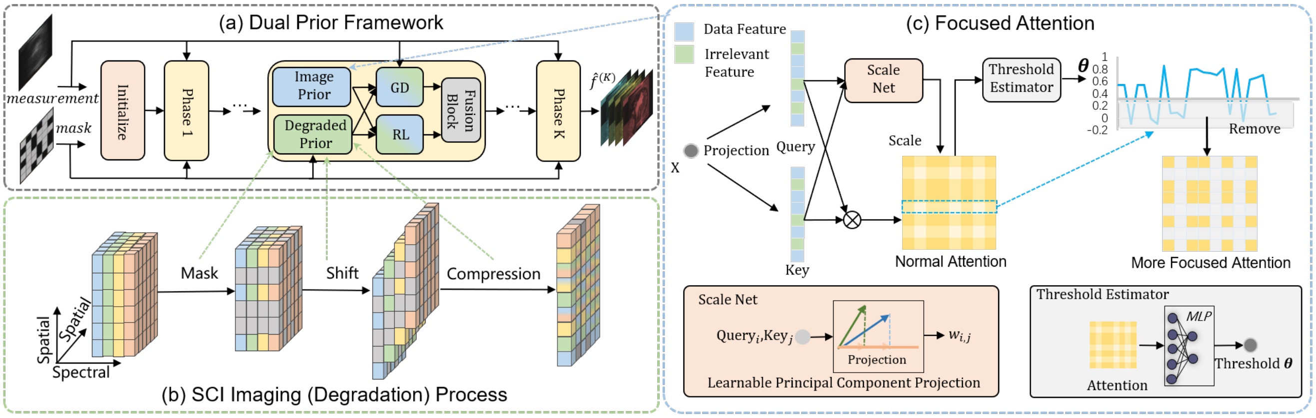

Figure 2. **Illustration of our main idea.** Dual Prior Framework (DPF): Multiple degradations are taken into account to formulate a degraded prior, which is subsequently integrated with the image prior through a combination of gradient descent (GD) and residual learning (RL). This fusion enables the simultaneous utilization of the two priors, thereby facilitating dual reconstruction within a single iteration. Focused Attention (FA): Leveraging inspiration from PCA denoising, we employ a learnable principal component projection to scale self-attention. Subsequently, we utilize thresholds to effectively eliminate irrelevant features from self-attention, enhancing the transformer's reconstruction capabilities. By incorporating FA as the image prior within the DPF, our Dual Prior Unfolding method is formulated.

veloped the Dual Prior Unfolding (DPU) method. This method aims to jointly harness two or more deep priors while significantly enhancing the iteration efficiency of unfolding methods. Our unfolding method comprises two crucial components: the Dual Prior Framework (DPF) and Focused Attention (FA). DPF, beyond the typical image prior, introduces a residual into the iteration formula to establish a new reconstruction prior, i.e., the degraded prior for the residual. This novel prior accounts for various degradation aspects, forming the fundamental framework. Further refinement is achieved by integrating a formula-free framework based on residual learning [17]. As shown in Fig. 2 (a) and (b), considering the imaging process as a degradation sequence, at the k-th iteration, DPF performs simultaneous preliminary restoration of the degraded image based on the image prior. Additionally, it estimates the degraded residual guided by the newly constructed degraded prior. These components are integrated via gradient descent and residual learning, resulting in dual output. The final output is achieved via the fusion block, allowing efficient utilization of multiple deep priors and enhancing iteration efficiency.

Moreover, we introduce the Focused Attention (FA) mechanism to optimize the reconstruction efficiency of the image prior as demonstrated in Fig. 2 (c). FA stands as a tailored enhancement technique for self-attention mechanisms, integrating a Scale Net utilizing a learnable principal component projection. This Scale Net adjusts attention size, accentuating crucial features while mitigating noisy ones. Additionally, a Threshold Net is implemented to efficiently filter out irrelevant features. We leverage a shifted window (Swin) attention approach to capture non-local similarity within the HSI. In efforts to curtail computational costs, we engineer an asymmetric backbone structure based on the U-Net architecture, specifically designed for hierarchical models like the Swin Transformer [25]. This adaptation results in a notable reduction—halving both computational requirements and parameter count within the transformers. The contributions of our work are as follows:

- We introduce a Dual Prior Unfolding SCI reconstruction model, which achieves the joint utilization of multiple deep priors and greatly improves iteration efficiency.
- We present a versatile Focused Attention mechanism as the image prior for DPF in our DPU framework. This approach directs the network's attention towards more pertinent features and can be extended to general tasks.
- To decrease the computational overhead of the fundamental transformer architecture while maintaining its hierarchical characteristics, we propose an asymmetric backbone. This modification is also applicable and beneficial for other hierarchical network architectures.
- Our approach demonstrates high performance with clear results, achieved with minimal computational and memory costs in both simulation and real-world experiments.

2. Realated Work

2.1. Model-based and Learning-based methods

SCI reconstruction approaches fall into two broad categories: model-based and learning-based. Initially, model-based methods [1, 23, 24, 36, 42, 46] employ hand-crafted priors (e.g., low rank [24, 51], sparsity [20, 36], total variation) to formulate optimization problems, solved iteratively. However, these methods are often inefficient and struggle to yield satisfactory results. As deep learning has made remarkable achievements in other fields, such as object detection[8, 12, 35, 37, 43], image restoration [7, 22, 53], image classification[15, 17] and so on, some learning-based methods [3, 4, 6, 18, 28, 33, 44, 45, 54] have been proposed to learn the mapping between degraded images and reconstructed images. Although the problem of reconstruction speed is solved, the reconstruction results are still unsatis-

Framework	GAP [32] IJCV 2023		DAUF [5] NeurIPS 2022		RDLF [14] CVPR 2023		DPF Ours
stage	5stg	9stg	5stg	9stg	5stg	9stg	5stg
PSNR	38.39	39.30	38.63	39.55	38.84	39.60	**39.62**
SSIM	0.965	0.971	0.965	0.972	0.969	**0.973**	**0.973**
Params (M)	**1.51**	2.71	1.55	2.76	1.76	3.17	1.59
FLOPs (G)	**22.77**	40.90	23.52	41.71	28.18	51.11	27.41

Table 1. Comparison of the proposed DPF and SOTA Unfolding Frameworks. The DPF achieves the best 5-stage unfolding performance even better than other frameworks' 9-stage performance, which demonstrates the effectiveness and efficiency of our DPF.

factory, and the methods are not interpretable.

With these problems, plug-and-play (PnP) methods and unfolding methods offer different solutions with interpretability. PnP methods [31, 47, 56] replace the handcraft priors with pre-trained networks as denoising priors. Although it achieves better results than traditional model-based methods, it is still limited by iteration efficiency. In contrast, deep unfolding methods [19, 27, 40, 41, 52] take deep networks as the learnable priors of optimization algorithms, which can achieve higher reconstruction quality through end-to-end training and learning while reducing the number of iterations to several times.

2.2. Deep Unfolding Methods

The deep unfolding method, combining model-based and learning-based approaches, shows promise for SCI reconstruction. DGSMP [19] uses an iterative framework with MAP estimation and a learnable Gaussian Scale Mixture prior. HerosNet [52] improves inter-stage interaction and parameter adaptation. DAUHST [5] addresses degradation patterns and ill-posedness with a degradation-aware framework and half-shuffle attention. RDLUF [14] jointly exploits spatial and spectral priors, and estimates the sensing matrix using degradation information. Despite successes, the growing computational demands pose challenges for deep unfolding methods. Besides, we can intuitively see the development trend of considering more prior information to achieve higher performance in these unfolding methods. However, the traditional unfolding framework with a single prior limits the effective utilization of more priors and the efficiency of the unfolding methods.

3. Proposed Method

3.1. Degradation of Snapshot Compressive Imaging

The coded aperture snapshot spectral compressive imaging (CASSI) system is the most popular SCI system at present, and the imaging process is shown in Fig. 2(b). Mathematically, we assume a spectral image patch with Λ bands $\{F_i\}_{i=1}^{\Lambda} \in \mathbb{R}^{H\times W}$, image frame F_λ is modulated by a physical mask with pattern $M \in \mathbb{R}^{H\times W}$. Then the modulated image frames of different wavelengths are shifted spatially and summed element-wise. Therefore, the modulated HSI frames $\{F_i\}_{i=1}^{\Lambda}$ are compressed to a coded measurement $G \in \mathbb{R}^{H\times(W+d(\Lambda-1))}$:

$$G\left(m,n\right) = \sum_{i=1}^{\Lambda} M \odot F_i\left(m, n+d\left(i-1\right)\right), \tag{1}$$

where $\odot$ means element-wise (Hadamard) product, m and n index the spatial coordinates, d represents pixels shift between adjacent bands. The SCI model represented in Eq. (1) can be expressed in a matrix-vector format as $g = \Phi f$, wherein $g \in \mathbb{R}^{H(W+d(\Lambda-1))}$ and $f \in \mathbb{R}^{H(W+d(\Lambda-1))\Lambda}$ denote the vectorized representations of the compressive image G and the original spectral image F, respectively. Moreover, $\Phi \in \mathbb{R}^{H(W+d(\Lambda-1))\times H(W+d(\Lambda-1))\Lambda}$ serves as the sensing matrix. While previous methods acknowledged prior information regarding mask, introducing mechanisms like mask guidance [3], degradation-aware techniques [5], and degradation learning approaches [14], they often overlooked degradation induced by shift and compression.

3.2. Dual Prior Framework

Previous unfolding methods [5, 14, 19, 52] have predominantly addressed limited degradation issues within the constraints of a single image prior, posing challenges in effectively leveraging multiple priors and impeding iteration efficiency. To address this limitation, we introduce a Dual Prior Framework (DPF), as illustrated in Fig. 2 (a). This framework accommodates increased degradation considerations, allowing the construction of a degraded prior. By doing so, it enables the efficient utilization of two or even more deep priors, enhancing iteration efficiency. The resultant deep unfolding implementation is obtained by introducing a residual and optimizing the following problem:

$$\arg\min_{f,z,r} \frac{1}{2}\|g-\Phi f\|^2 + \gamma D(z) + \tau R(r), s.t., f = z - r, \tag{2}$$

where $\|g-\Phi f\|^2$ is the data fidelity term; $z \in \mathbb{R}^{H(W+d(\Lambda-1))\Lambda}$ is the preliminary restored image; $r \in \mathbb{R}^{H(W+d(\Lambda-1))\Lambda}$ is the residual associated with the degradation pattern; $D(\cdot)$ represents the image prior; $R(\cdot)$ is a degraded prior, and γ, τ are tradeoff parameters.

We adopt the Augmented Lagrange Method (ALM) for its accuracy and fast convergence to obtain an unfolding inference. Then Eq. (2) is changed into the Augmented Lagrange formulation as

$$\begin{aligned} L(f,z,r,y,\mu) = \frac{1}{2}\|g-\Phi f\|^2 + \frac{\mu}{2}\|f - z + r + \frac{y}{\mu}\|^2 \\ + \gamma D(z) + \tau R(r), \end{aligned} \tag{3}$$

where y is the Lagrange multiplier and μ is the penalty parameter. Subsequently, Eq. (3) can be solved by alternately updating r, z, f. For r and z sub-problems, they are a particular case of the so-called proximal mapping, i.e., $prox_{\lambda h}(x)$ as follows:

$$prox_{\lambda h}(x) = \arg\min_{x} \frac{1}{2}\|x-s\|^2 + \lambda h(x), \tag{4}$$

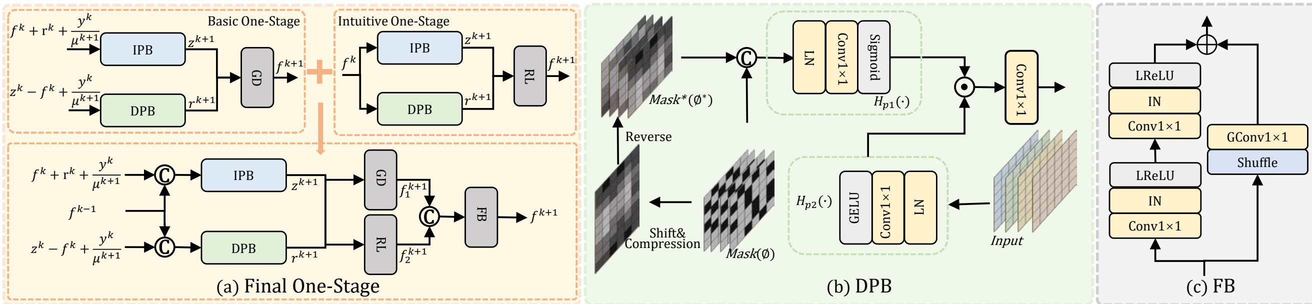

Figure 3. Details of DPF one-stages. (a) Three different one-stage constructions. (b) and (c) The components of DPB and FB.

where $\lambda = \frac{\tau}{\mu^{k+1}}$, $h(x) = R(r)$, $s = z^k - f^k + \frac{y^k}{\mu^{k+1}}$ for r sub-problem, and $\lambda = \frac{\gamma}{\mu^{k+1}}$, $h(x) = D(z)$, $s = f^k + r^k + \frac{y^k}{\mu^{k+1}}$ for z sub-problem. To solve the r sub-problem, we construct a Degraded Prior Block (DPB) to learn mapping functions from mask, shift, and compression degradation, as illustrated in Fig. 3(b). The reverse operation in Fig. 3(b) is proposed in MST initialization [3] and defined as follows:

$$X_i\left(m,n\right) = 2G\left(m, n - d\left(i-1\right)\right)/\Lambda, \tag{5}$$

where $\{X_i\}_{i=1}^{\Lambda} \in \mathbb{R}^{H\times W}$ is the reconstructed data patch with Λ channels, $G \in \mathbb{R}^{H\times(W+d(i-1))}$ represents a shift and compressed measurement. To utilize compression and shift degradation, we first shift, compress, and reverse the mask to get a new mask that contains various degradation. Then DPB can learn these degradations from the differences between the new mask and the original mask through a 1×1 convolution kernel ($conv1\times1$), and return an element-wise degradation weight via the Sigmoid activation function to filter the feature. As demonstrated in Fig. 3(b), we estimate the proximal mapping as follows:

$$r^{k+1} = H_{p1}(\Phi, \Phi^*) \odot H_{p2}(z^k - f^k + \frac{y^k}{\mu^{k+1}}), \tag{6}$$

where $\odot$ denotes element-wise product; Φ^* is new mask in Fig. 3(b). We found that the prior networks have adaptive adjustment ability and the tradeoff parameters $\frac{\tau}{\mu^{k+1}}$ and $\frac{\gamma}{\mu^{k+1}}$ have little effect on the reconstruction, so we ignore the tradeoff parameters in prior networks. Note that we will depict that the z sub-problem in Eq. (3) can be solved with a transformer-based focused attention in Sec. 3.3. Here, we skip the details about the z sub-problem and give a general solver to enable the subsequent deduction:

$$z^{k+1} = IPB(f^k + r^k + \frac{y^k}{\mu^{k+1}}). \tag{7}$$

The data fidelity term within Eq. (3) is associated with a quadratic regularized least-squares problem as follows:

$$f^{k+1} = \arg\min_f \|g - \Phi f\|^2 + \mu^{k+1}\|(z^{k+1} - r^{k+1} - \frac{y^k}{\mu^{k+1}}) - f\|^2. \tag{8}$$

Considering the form of the sensing matrix Φ and the Sherman-Morrison-Woodbury matrix inversion lemma, the above formula has a closed-form solution[5, 41] as follows:

$$f^{k+1} = z^{k+1} - r^{k+1} - \frac{y^k}{\mu^{k+1}} + \Phi^T \frac{g - \Phi(z^{k+1} - r^{k+1} - \frac{y^k}{\mu^{k+1}})}{\mu^{k+1} + \Phi\Phi^T}. \tag{9}$$

Eq. (9) is a special form of Gradient Descent (GD), so we can get the basic one-stage of DPF as shown in the top-left of Fig. 3(a). Besides, we also propose an intuitive one-stage based on the Residual Learning (RL) strategy [17, 50], which is a complement to the basic scheme. As illustrated in the top-right of Fig. 3(a), f^k is inputted to IPB and DPB respectively to get the initial recovery image z^{k+1} and residual r^{k+1}, and then subtract r^{k+1} from z^{k+1} to get the final output f^{k+1}, which is a formula-free residual learning network. Combining this intuitive one-stage with the basic one-stage, our final one-stage of DPF is established as demonstrated at the bottom of Fig. 3(a). Among them, a Fusion Block (FB) is proposed to fuse the results of the two and get the final output, which is detailed in Fig. 3(c).

Finally, r^0, y^0 is initialized to 0, z^0 and f^0 are equally initialized by reversing the measurement and then embedding mask information with $conv1\times1$, and the Lagrange multiplier y^{k+1} is updated as:

$$y^{k+1} = y^k + \mu^{k+1}(f^{k+1} - z^{k+1} + r^{k+1}). \tag{10}$$

3.3. Focused Transformer&Attention

Inspired by Swin Transformer's success in visual tasks [21, 25], we introduced the Swin transformer to capture non-local spatial similarities in SCI reconstruction. However, as a hierarchical network, Swin Transformer requires twice the number of parameters and computations compared to Transformers such as MST [3] and DAUHST [5]. To improve efficiency and maintain hierarchical characteristics, we propose an asymmetric backbone for hierarchical networks. To further exert the reconstruction capability of self-attention, we propose a Focused Attention (FA) inspired by PCA denoising, which adopts learnable principal component projection and threshold network to make the network pay more attention to important features and improve reconstruction efficiency. Finally, we insert FA into the Swin Transformer based on an asymmetric backbone to build our Focused Transformer (FT), which is also our IPB.

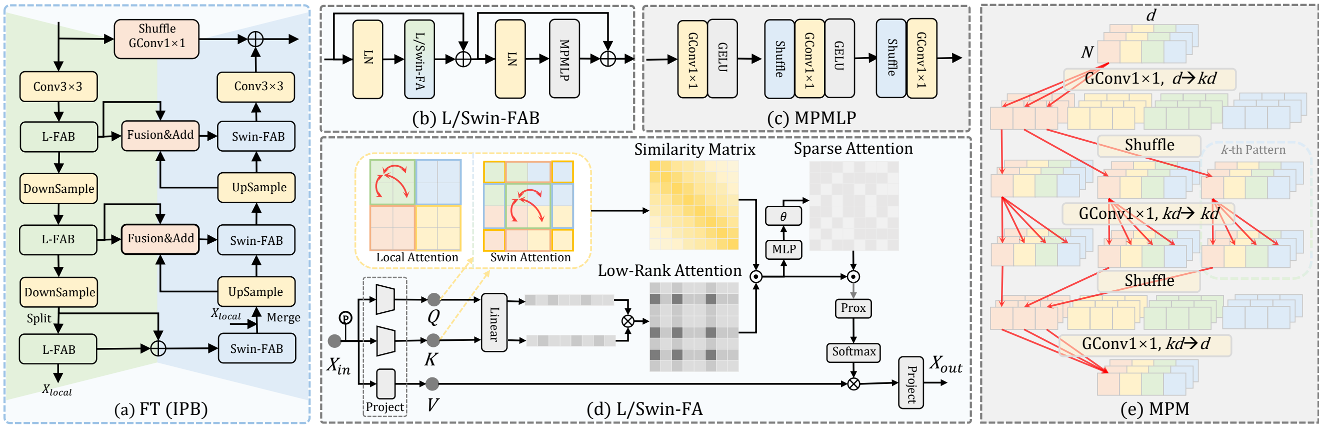

Figure 4. Details of the FT and some critical components of transformer blocks. (a) The backbone structure of FT. (b) and (c) The components of L/Swin-FAB and MPMLP. (d) Details of L/Swin-FA. (e) Illustration of Multi-Pattern Mechanism.

Network Architecture. As shown in Fig. 4(a), FT adopts a three-level asymmetric Unet backbone built by the basic unit Local/Swin Focused Attention Block (L/Swin-FAB). Firstly, $conv3 \times 3$ is adopted to extract and enhance features in both input and output. Each encoder or decoder layer contains a L/Swin-FAB and a resizing module. At the bottom layer, the downsampling features are split channel-wise, with one half fed into L-FAB and the other into Swin-FB, then the two are merged into the upsampling module. In Fig.4(b), L/Swin-FAB consists of two Layer Normalization (LN), a L/Swin-FA, and a Multi-Pattern Multilayer Perception (MPMLP) that is detailed in Fig. 4(c)and(e). We adopt $conv1 \times 1$ to fuse skip connection with upsampling features, followed by a residual connection. The downsampling and upsampling modules are stridden $conv4 \times 4$ and $deconv2 \times 2$. Finally, shuffle and $gconv1 \times 1$ operate on input, which is added to the features obtained through Unet to obtain the output.

Asymmetric Backbone for Hierarchical Network. When many hierarchical networks e.g. Swin Transformer [25] and local-global transformers [13, 39], there is an effective non-local information modeling capability but also some computational burdens. To solve this problem, the proposed asymmetric backbone utilizes the skip connection of Unet to reduce computation without destroying the properties of the hierarchy, as shown in Fig. 4(a). Specifically, we divide the UNet backbone into two parts, where the left part calculates attention between pixels in the local windows and the right part calculates attention between pixels in the shifted windows in our model. Then the outputs from the left part are transported to the right part as inputs through the skip connection, which achieves the hierarchical interaction between the two parts. For a general hierarchical network, the left and right (even the middle) adopt different modules, and the skip connection achieves hierarchy preservation while halving the computation and memory cost.

Low-Rank Attention based on Principle Component Projection. In PCA denoising, we could calculate the maximum energy projection direction of the data distribution through statistics. Inspired by this, we set up a learnable principal component with the same dimension as the feature and trained it to learn the distribution of the data. Then the feature projection on the principal component is used to scale the matching attention so that the features that conform to the trend of data distribution get more attention. Specifically, the input tokens $X_{in} \in R^{HW \times C}$ embedded by an implicit position [11] are denoted as X_{pe}. Subsequently, X_{pe} is linearly projected into $query\ Q \in R^{HW \times \Lambda}$, $key\ K \in R^{HW \times \Lambda}$ and $value\ V \in R^{HW \times C}$ as

$$Q = X_{pe} W_Q,\ K = X_{pe} W_K,\ V = X_{pe} W_V, \tag{11}$$

where $W_Q, W_K \in R^{C \times \Lambda}$ and $W_V \in R^{C \times C}$ are learnable parameters and Λ is the number of bands in HSIs. When the channels of the feature are essentially extended from the bands of the HSIs, we project all Q and K to the Λ dimension to reduce the computation. For simplicity, we will only use the notation C in the following sections regardless of the number of channels. As shown in Fig. 4(d), Q, K, V are partitioned into non-overlapping windows of $B \times B$ tokens and reshaped into $R^{\frac{HW}{B^2} \times B^2 \times C}$. Subsequently, Q, K, V are split along the channels wise into N heads: $Q = [Q_1, Q_2, \cdots, Q_N]$, $K = [K_1, K_2, \cdots, K_N]$, and $V = [V_1, V_2, \cdots, V_N]$ and the dimension of each head is $d = \frac{C}{N}$. The self-attention similarity matrix M_i is calculated inside each head as Eq. (12)

Then, Q and K are used to compute the initial similarity matrix M and the vector q and k containing scaling factors for tokens in Q and K as follows:

$$M_i = Q_i K_i^T,\ q_i = Q_i W_q,\ k_i = K_i W_k, \tag{12}$$

where $W_q, W_k \in R^{\frac{C}{N} \times 1}$ are learnable principle components and biases are omitted for simplification. To simplify the scaling process, q and k are used to compute a scaling matrix $atten_{LR}$ and the scaled similarity matrix $M^{'}$ is further obtained as follows:

$$M_i^{'} = atten_{LR_i} \odot M_i,\ atten_{LRi} = q_i k_i^T, \tag{13}$$

Method	TSA-Net [29]		DGSMP [19]		GAP-Net [32]		HDNet [18]		MST [3]		CST [2]		BIRNAT [9]		DAUHST-9stg [5]		RDLUF [14]		DPU-5stg		DPU-9stg	
Category	CNN		Deep Unfolding		Deep Unfolding		CNN		Transformer		Transformer		Recurrent CNN		Deep Unfolding		Deep Unfolding		Deep Unfolding		Deep Unfolding	
Reference	ECCV 2020		CVPR 2021		IJCV 2023		CVPR 2022		CVPR 2022		ECCV 2022		TPMAI 2023		NeurIPS 2022		CVPR 2023		Ours		Ours	
Scene	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
1	32.03	0.892	33.26	0.915	33.74	0.911	35.14	0.935	35.40	0.941	35.96	0.949	36.79	0.951	37.25	0.958	37.94	0.966	38.19	0.964	**38.91**	**0.968**
2	31.00	0.858	32.09	0.898	33.26	0.900	35.67	0.940	35.87	0.944	36.84	0.955	37.89	0.957	39.02	0.967	40.95	0.977	40.57	0.975	**41.99**	**0.981**
3	32.25	0.915	33.06	0.925	34.28	0.929	36.03	0.943	36.51	0.953	38.16	0.962	40.61	0.971	41.05	0.971	43.25	0.979	43.11	0.977	**44.10**	**0.980**
4	39.19	0.953	40.54	0.964	41.03	0.967	42.30	0.969	42.27	0.973	42.44	0.975	46.94	0.985	46.15	0.983	47.83	**0.990**	47.78	0.988	**48.33**	**0.990**
5	29.39	0.884	28.86	0.882	31.44	0.919	32.69	0.946	32.77	0.947	33.25	0.955	35.42	0.964	35.80	0.969	37.11	0.976	37.43	0.975	**38.07**	**0.978**
6	31.44	0.908	33.08	0.937	32.40	0.925	34.46	0.952	34.80	0.955	35.72	0.963	35.30	0.959	37.08	0.970	37.47	0.975	37.49	0.973	**38.58**	**0.978**
7	30.32	0.878	30.74	0.886	32.27	0.902	33.67	0.926	33.66	0.925	34.86	0.944	36.58	0.955	37.57	0.963	38.58	0.969	38.17	0.967	**39.13**	**0.971**
8	29.35	0.888	31.55	0.923	30.46	0.905	32.48	0.941	32.67	0.948	34.34	0.961	33.96	0.956	35.10	0.966	35.50	0.970	36.13	0.970	**36.90**	**0.975**
9	30.01	0.890	31.66	0.911	33.51	0.915	34.89	0.942	35.39	0.949	36.51	0.957	39.47	0.970	40.02	0.970	41.83	0.978	41.77	0.977	**42.88**	**0.981**
10	29.59	0.874	31.44	0.925	30.24	0.895	32.38	0.937	32.50	0.941	32.09	0.945	32.80	0.938	34.59	0.956	35.23	0.962	35.55	0.964	**36.36**	**0.970**
Avg	31.46	0.894	32.63	0.917	33.26	0.917	34.97	0.943	35.18	0.948	36.12	0.957	37.58	0.960	38.36	0.967	39.57	0.974	39.62	0.973	**40.52**	**0.977**
Params	42.20M		3.58M (0.90M)		4.27M (0.47M)		2.25M		2.03M		3.00M		4.40M		6.15M (0.68M)		1.81M (0.60M)		**1.59M (0.31M)**		2.85M (**0.31M**)	
GFLOPs	125.75		84.77 (21.19)		78.58 (8.75)		154.76		28.15		40.10		2122.66		79.50 (9.9)		115.16 (12.80)		**27.41 (5.48)**		49.26 (**5.48**)	

Table 2. Comparisons between DPU and SOTA methods on 10 simulation scenes. PSNR in dB (left entry in each cell), SSIM (right entry in each cell), Params, and FLOPs are reported for all methods and the additional single-stage Memory and FLOPs are reported for unfolding methods. The best results are highlighted in bold.

where $atten_{LRi} \in R^{\frac{HW}{B^2} \times \frac{HW}{B^2}}$ is the Low-Rank Attention corresponding to the similarity matrix element-wisely.

Sparse Attention based on Threshold Filtering. In traditional signal processing, it is common to remove noisy components by projecting the signal into a specific space and eliminating small components that usually represent noise, such as the Principal Component Analysis (PCA) method for noise removal, the proximal mapping for Eq. (6) when $h(x) = \|\Psi(x)\|_1$ and Ψ is the projection operator [49]. Inspired by this, we take the calculated similarity matrix as a particular projection space and eliminate unimportant and irrelevant attention through threshold filtering. It is noted that different from traditional proximal mapping, activation value 0 in the similarity matrix still has much influence after passing through the $softmax$ activation function, so we need to take $-\infty$ to remove the irrelevant term. We define the particular proximal mapping as follows:

$$prox(M) = \begin{cases} M, & M > \theta, \\ -\infty, & M \leq \theta, \end{cases} \quad \theta = H_\theta(M - D), \tag{14}$$

where θ is a threshold estimated through a Multilayer Perception (MLP) $H_\theta(\cdot)$ consisting of linear layer and $LReLU$, and D is a diagonal matrix whose diagonal elements are the diagonals of M. Since the self-similarity of each token will interfere with the threshold estimation, we remove the diagonal element in Eq. (14). Here we present two schemes based on sparse index and threshold operator to implement this proximal mapping: as shown in the right of Fig. 4(d), one is that the noise in attention is set to 0 by sparse attention and then passes through the $prox$ with a threshold of 0, which achieves the following effect:

$$Atten_i = softmax(prox(M_i^{'}))V, \tag{15}$$

the other is directly applied to the final self-attention,

$$\begin{aligned} &Atten_i = (softmax(M_i^{'}) \odot atten_{Si})V, \\ &atten_{Si} = (M_i^{'} > \theta_i), \end{aligned} \tag{16}$$

where $Atten_i$ is final self-attention inside each head; $atten_{Si} \in R^{\frac{HW}{B^2} \times \frac{HW}{B^2}}$ is the Sparse Attention composed of 0, 1 by threshold estimation. Finally, the outputs of N heads are concatenated in channel-wise, reshape into $R^{HW \times C}$ to undergo a linear projection:

$$X_{out} = concat_i^N(Atten_i)W + b, \tag{17}$$

where $X_{out} \in R^{HW \times C}$ is the final output; $W \in R^{C \times C}$ is learnable parameters and b is a learnable bias.

Multi-Pattern Multilayer Perception (MPMLP). Following classic vision transformers design [15, 25], we take an MLP after self-attention to mix spectral (channel) information. However, normal fully connected MLP can be quite burdensome, thus we propose an MPMLP inspired by the multi-head self-attention to further reduce the number of parameters and computation costs, as shown in Fig. 4(c). To have a better understanding of MPMLP, we demonstrate the Multi-Pattern Mechanism (MPM) in Fig. 4(e).

4. Experiment

4.1. Datasets and Evaluations

Datasets. We evaluate our DPU method on both simulation and real datasets. The simulation experiments are conducted on the public HSI datasets CAVE [34] and KAIST [10]. Following the settings of TSA-Net [29], we adopt the real mask of size 256 $\times$ 256 for simulation. The CAVE dataset is used to train the network and 10 scenes with the spatial size of 256 $\times$ 256 are extracted from the KAIST dataset for testing. For the experiments on the real scenes, 5 real HSI compressive measurements with a spatial size of 660 $\times$ 714 captured by the CASSI system developed in TSA-Net [29] are utilized for testing.

Comparison Methods: We compare our DPU method on synthetic data with SOTA reconstruction methods including HDNet [18], TSA-Net [29], BIRNAT [9], and unfold-

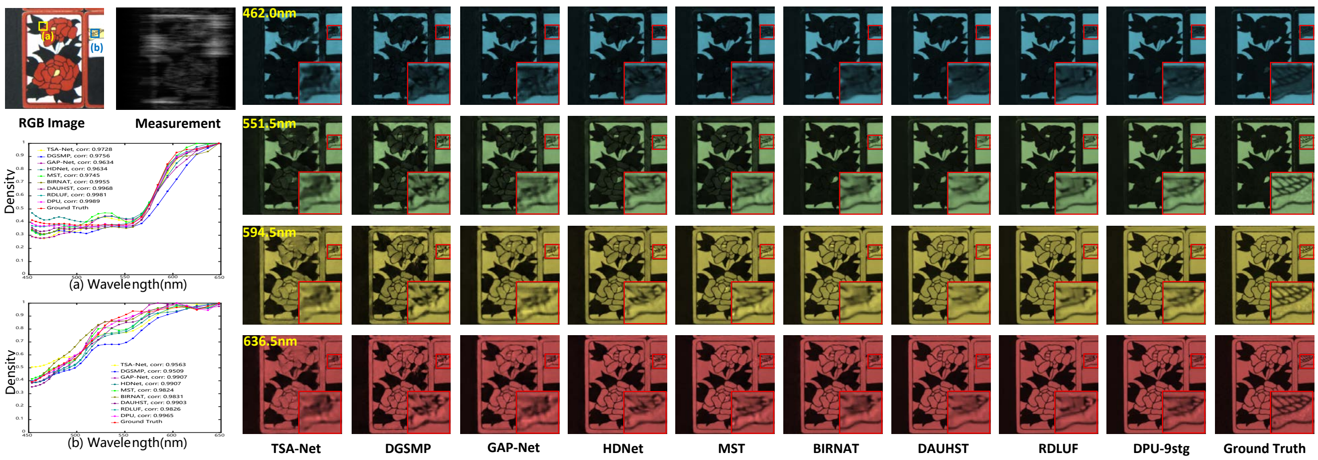

Figure 5. Reconstructed images of simulation scene 7 with 4 out of 28 spectral channels by the state-of-the-art methods. Two regions in scene 7 are selected for analyzing the spectra of the reconstructed results. The figure is better viewed in a zoomed-in PDF.

ing methods: DGSMP [19], GAP-Net [32], DAUHST [5], RDLUF [14] and Transformer methods: MST [3], CST [2].

Implementation Details. Our DPU is implemented in PyTorch and trained using a single RTX3090 GPU. We adopt the multi-stage root mean square error (RMSE) loss function [55] and Adam optimizer with setting $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 10^{-8}$ to train the proposed network. We take the similar network in [5] to estimate the hyperparameters. The window size of basic Swin Attention is set to 8×8. We set the initial learning rate as 4×10^{-4} and adopt the Cosine Annealing learning rate scheme [26] to implement end-to-end training. Following most previous unfolding methods, we set the maximum number of iterations to 9, i.e., DPU-9stg. Finally, more content and results are provided in the supplementary materials to have a better understanding.

4.2. Quantitative Results

As shown in Table 2, we compare the PSNR, SSIM, Memory, and FLOPs of DPU and SOTA methods. To intuitively show the effectiveness of our DPU, we provide Performance-FLOPs-Params comparisons of SOTA methods in Fig. 1. The proposed DPU obtains 40.52dB of PSNR and 0.977 of SSIM, outperforming competing methods. To be noted, compared with the SOTA method RDLUF, our DPU-9stg achieves 0.95dB/0.003 improvement on PSNR/SSIM with less than $1/2$ FLOPs and single-stage memory, and DPU-5stg achieves better performance with less than $1/4$ FLOPs and less memory. Compared with the SOTA RNN method BIRNAT, the proposed DPU-5stg achieves a 2.04dB improvement of PSNR and 0.013 improvement of SSIM while only requiring about $1/2$ parameters and $1/77$ FLOPs in Table. 2. Finally, our DPU-5stg outperforms other methods with the least FLOPs and parameters when DPU-9stg significantly outperforms other unfolding methods with the least single-stage FLOPs and parameters, which demonstrates the reconstruction effectiveness and efficiency of DPU.

4.3. Qualitative Results

Simulation Data Results. As shown in Fig. 5, the reconstructed HSIs produced by the DPU restore more sharp edge textures and fewer undesirable artifacts in different spectral channels than other competing methods. In the comparisons of spectral curves, the DPU has the highest correlation and the most similar shape to the ground truth. In addition, the proposed DPU provides clearer pattern details, sharper line outlines, and less blurry deformation, while the results of the other unfolding methods are blurry to some extent, which also shows the efficacy of our method.

Real Data Results. We also apply our DPU to address the real-scene HSI reconstruction. In the experiment, we train DPU with the real mask on the CAVE and KAIST datasets jointly under the same settings as [5, 19, 29]. 11-bit shot noise is also added into the measurements during training to simulate the real degradation and the visual comparisons with SOTA methods are shown in Fig. 6. Intuitively, our DPU obtains better visualization with a smooth texture and clear details while other methods produce more distortion and blurred details. In the last two of the four bands, we can even see the strawberry seeds clearly, which is difficult for other methods. This evidence proves the powerful reconstruction ability of DPU and suggests that DPU is more robust and practical for real-scene HSI reconstruction.

4.4. Ablation Study

To assess the individual contributions of various components within the proposed DPU framework, as well as the efficacy of the Degraded Prior Fusion (DPF) and transformer modules, we undertake a series of ablation studies on both the CAVE and KAIST datasets.

Break-down Ablation. We adopt baseline-1, which is derived by retaining the base iteration formula and removing L/Swin-FA from DPU-5stg to conduct the breakdown ablation, to study the effect of each principal component. Table 3 shows the results of PSNR and SSIM on different settings and baseline-1 yields 37.28dB. The model achieves

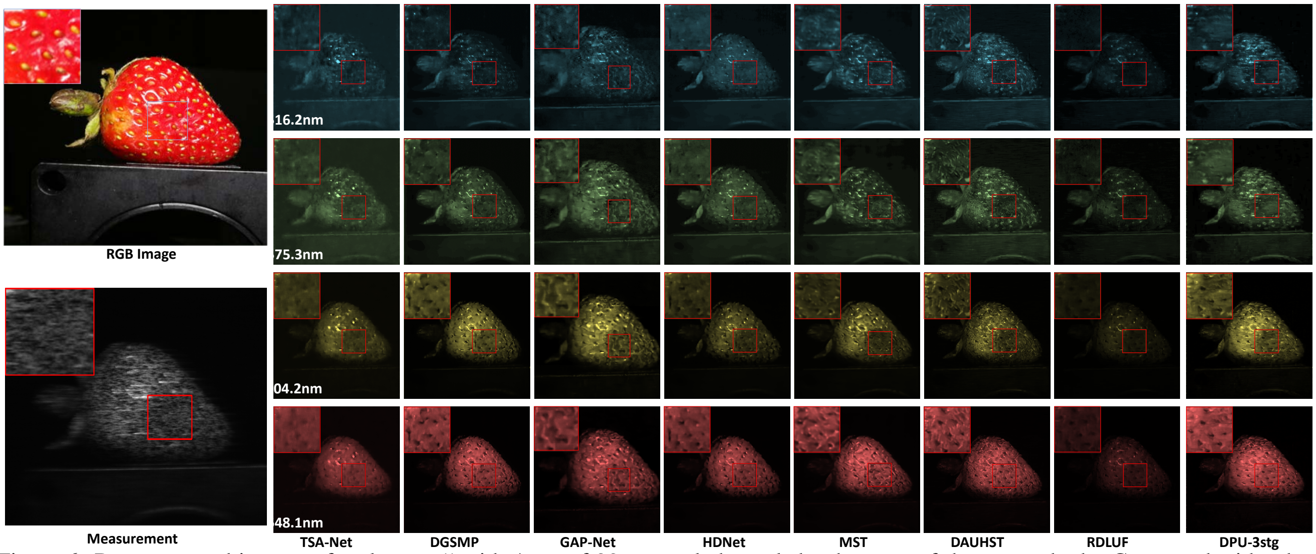

Figure 6. Reconstructed images of real scene 5 with 4 out of 28 spectral channels by the state-of-the-art methods. Compared with other methods, our DPU recovers more details and clear content.

Method	Base line-1	L/Swin-FA	+Intuitive DPF	+Basic DPF	+DPF
PSNR	37.28	38.49	38.76	39.23	39.62
SSIM	0.958	0.966	0.968	0.971	0.973
Params (M)	1.15	1.52	1.55	1.55	1.59
FLOPs (G)	15.79	23.05	24.95	24.95	27.41

Table 3. Break-down ablation study.

Method	Base line-2	S-MSA [3]	Swin(H) [25]	HS-MSA [5]	Swin*	Swin*+FA
PSNR	34.78	35.62	35.97	36.08	36.27	36.58
SSIM	0.938	0.950	0.952	0.952	0.953	0.954
Params (M)	1.14	1.68	1.68	1.68	1.38	1.51
FLOPs (G)	14.81	22.16	24.46	26.10	22.57	22.77

Table 4. Attention comparison. Swin* represents Swin-MSA [25] based on our asymmetric backbone.

1.21 and 1.13dB improvements when we successively apply L/Swin-FA and DPF. In addition, we further investigate the two frameworks in DPF, and the results show that the framework derived from the basic formula has a higher gain of 0.74dB, while the intuitive framework can supplement the additional gain 0.39dB. These results demonstrate the effectiveness of our L/Swin-FA and DPF.

Unfolding Framework Comparison. Table. 1 reports the results of the comparison of unfolding frameworks. All frameworks are implemented on our DPU, DPF achieves an obvious improvement of 1.23, 0.99, and 0.78dB higher than SOTA framework GAP [32], DAUF [5], and RDLF [14] in the 5-stage unfolding. In addition, we further measured the performance of other frameworks in the 9-stage unfolding, and the results show that our 5-stage approach achieves better performance than other 9-stage frameworks with less computation and parameters, which intuitively demonstrates the efficiency and effectiveness of DPF.

Attention Comparison. To study the effect of transformer modules, we perform ablation of L/Swin-FA and other self-attention. Baseline-2 is obtained by removing L/Swin-FA and the iterative formula from DPU-5stg. S-MSA[3], Swin(H)[25] and HS-MSA[5] use the original unet backbone when L/Swin* and L/Swin*+FA adopt our asymmetric backbone. Swin(H) is implemented using the half operation of HS-MSA [5] for a fair comparison. As shown in Table 4, baseline-2 yields 34.78dB. L/Swin*+FA yields the most significant improvement of 1.8dB, 0.96, 0.61, and 0.5dB higher than S-MSA, Swin(H), and HS-MSA with almost the least parameters and FLOPs. When we exploit Swin* and FA successively, 0.3dB and 0.61dB gains than Swin(H) are achieved when require less Memory and computation, which demonstrates the effectiveness of asymmetric backbone and focused attention.

5. Conclusion

This study introduces an effective and efficient deep unfolding approach, denoted as DPU, specifically designed for hyperspectral SCI reconstruction. The DPU method is initially structured by a novel dual prior framework, strategically incorporating focused attention within an iterative framework to improve reconstruction quality. This strategy efficiently harnesses the joint utilization of multiple priors while enhancing iteration efficiency. Moreover, an asymmetric backbone is devised to preserve hierarchical properties while simultaneously reducing computational requirements for the DPU method. Empirical validation through quantitative and ablation experiments substantiates the efficacy of the proposed approach.

Acknowledgements: This work was supported by the National Natural Science Foundation of China under Grant 62302394 and 62106063, the China Postdoctoral Science Foundation under Grant 317751, the Natural Science Foundation of Shaanxi under Grant 2023-JC-QN-0757, and the Shenzhen Science and Technology Program under Grant RCBS20210609103708013.

References

[1] José M. Bioucas-Dias and Mário A. T. Figueiredo. A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Transactions on Image Processing*, 16:2992–3004, 2007. 1, 2

[2] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 686–704. Springer, 2022. 6, 7

[3] Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17481–17490, 2022. 1, 2, 3, 4, 6, 7, 8

[4] Yuanhao Cai, Jing Lin, Zudi Lin, Haoqian Wang, Yulun Zhang, Hanspeter Pfister, Radu Timofte, and Luc Van Gool. Mst++: Multi-stage spectral-wise transformer for efficient spectral reconstruction. In *CVPRW*, 2022. 1, 2

[5] Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc Van Gool. Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging. In *Advances in Neural Information Processing Systems*, 2022. 1, 3, 4, 6, 7, 8

[6] Yuanhao Cai, Yuxin Zheng, Jing Lin, Xin Yuan, Yulun Zhang, and Haoqian Wang. Binarized spectral compressive imaging. In *Proc. Conf. Neural Inf. Process. Syst.*, 2023. 1, 2

[7] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22367–22377, 2023. 2

[8] Yihong Chen, Zheng Zhang, Yue Cao, Liwei Wang, Stephen Lin, and Han Hu. Reppoints v2: Verification meets regression for object detection. *Advances in Neural Information Processing Systems*, 33:5621–5631, 2020. 2

[9] 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, 2023. 6

[10] 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*, 36: 1–13, 2017. 6

[11] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional positional encodings for vision transformers. In *The Eleventh International Conference on Learning Representations*, 2022. 5

[12] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*, 29, 2016. 2

[13] Kexing Ding, Ting Lu, Wei Fu, Shutao Li, and Fuyan Ma. Global–local transformer network for hsi and lidar data joint classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. 5

[14] 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 (CVPR)*, pages 22262–22271, 2023. 1, 3, 6, 7, 8

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 6

[16] Michael E Gehm, Renu John, David J Brady, Rebecca M Willett, and Timothy J Schulz. Single-shot compressive spectral imaging with a dual-disperser architecture. *Optics express*, 15(21):14013–14027, 2007. 1

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 4

[18] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17521–17530, 2022. 1, 2, 6

[19] Tao Huang, Weisheng Dong, Xin Yuan, Jinjian Wu, and Guangming Shi. Deep gaussian scale mixture prior for spectral compressive imaging. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16211–16220, 2021. 1, 3, 6, 7

[20] 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. 2

[21] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 4

[22] Hezheng Lin, Xing Cheng, Xiangyu Wu, and Dong Shen. Cat: Cross attention in vision transformer. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2022. 2

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

[24] Yang Liu, Xin Yuan, Jinli Suo, David J. Brady, and Qionghai Dai. Rank minimization for snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:2990–3006, 2019. 1, 2

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer:

Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2, 4, 5, 6, 8

[26] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 7

[27] 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

[28] Ziyi Meng and Xin Yuan. Perception inspired deep neural networks for spectral snapshot compressive imaging. In *2021 IEEE International Conference on Image Processing*, pages 2813–2817, 2021. 1, 2

[29] Ziyi Meng, Jiawei Ma, and Xin Yuan. End-to-end low cost compressive spectral imaging with spatial-spectral self-attention. In *ECCV*, 2020. 6, 7

[30] Ziyi Meng, Mu Qiao, Jiawei Ma, Zhenming Yu, Kun Xu, and Xin Yuan. Snapshot multispectral endomicroscopy. *Optics Letters*, 45(14):3897–3900, 2020. 1

[31] Ziyi Meng, Zhenming Yu, Kun Xu, and Xin Yuan. Self-supervised neural networks for spectral snapshot compressive imaging. In *2021 IEEE/CVF International Conference on Computer Vision*, pages 2602–2611, 2021. 3

[32] Ziyi Meng, Xin Yuan, and Shirin Jalali. Deep unfolding for snapshot compressive imaging. *International Journal of Computer Vision*, 131(11):2933–2958, 2023. 1, 3, 6, 7, 8

[33] Xin Miao, Xin Yuan 0002, Yunchen Pu, and Vassilis Athitsos. lambda-net: Reconstruct hyperspectral images from a snapshot measurement. In *2019 IEEE/CVF International Conference on Computer Vision*, pages 4058–4068, 2019. 1, 2

[34] J. Park, M. Lee, M. D. Grossberg, and S. K. Nayar. Multispectral Imaging Using Multiplexed Illumination. In *IEEE International Conference on Computer Vision*, 2007. 6

[35] Christian Szegedy, Alexander Toshev, and Dumitru Erhan. Deep neural networks for object detection. *Advances in neural information processing systems*, 26, 2013. 2

[36] Jin Tan, Yanting Ma, Hoover F. Rueda, Dror Baron, and Gonzalo R. Arce. Compressive hyperspectral imaging via approximate message passing. *IEEE Journal of Selected Topics in Signal Processing*, 10:389–401, 2016. 1, 2

[37] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020. 2

[38] Ashwin A Wagadarikar, Nikos P Pitsianis, Xiaobai Sun, and David J Brady. Video rate spectral imaging using a coded aperture snapshot spectral imager. *Optics express*, 17(8): 6368–6388, 2009. 1

[39] Fengfeng Wang, Jie Li, Qiangqiang Yuan, and Liangpei Zhang. Local–global feature-aware transformer based residual network for hyperspectral image denoising. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022. 5

[40] Lizhi Wang, Chen Sun, Ying Fu, Min H. Kim, and Hua Huang. Hyperspectral image reconstruction using a deep spatial-spectral prior. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8024–8033, 2019. 3

[41] Lizhi Wang, Chen Sun, Maoqing Zhang, Ying Fu, and Hua Huang. Dnu: Deep non-local unrolling for computational spectral imaging. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1658–1668, 2020. 1, 3, 4

[42] Minghua Wang, Qiang Wang, and Jocelyn Chanussot. Tensor low-rank constraint and l_0 total variation for hyperspectral image mixed noise removal. *IEEE Journal of Selected Topics in Signal Processing*, 15:718–733, 2021. 1, 2

[43] Alexander Wong, Mahmoud Famuori, Mohammad Javad Shafiee, Francis Li, Brendan Chwyl, and Jonathan Chung. Yolo nano: A highly compact you only look once convolutional neural network for object detection. In *2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition (EMC2-NIPS)*, pages 22–25. IEEE, 2019. 2

[44] Zhiwei Xiong, Zhan Shi, Huiqun Li, Lizhi Wang, Dong Liu, and Feng Wu. Hscnn: Cnn-based hyperspectral image recovery from spectrally undersampled projections. *2017 IEEE International Conference on Computer Vision Workshops*, pages 518–525, 2017. 1, 2

[45] Kouhei Yorimoto and Xian-Hua Han. Hypermixnet: Hyperspectral image reconstruction with deep mixed network from a snapshot measurement. In *2021 IEEE/CVF International Conference on Computer Vision Workshops*, pages 1184–1193, 2021. 1, 2

[46] Xin Yuan. Generalized alternating projection based total variation minimization for compressive sensing. *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2539–2543, 2016. 1, 2

[47] Xin Yuan, Yang Liu, Jinli Suo, and Qionghai Dai. Plug-and-play algorithms for large-scale snapshot compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1447 – 1457, 2020. 3

[48] Xin Yuan, David J. Brady, and Aggelos K. Katsaggelos. Snapshot compressive imaging: Theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 38(2):65–88, 2021. 1

[49] Jian Zhang and Bernard Ghanem. ISTA-Net: Interpretable optimization-inspired deep network for image compressive sensing. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1828–1837, 2018. 6

[50] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on Image Processing*, 26(7):3142–3155, 2017. 4

[51] Shipeng Zhang, Lizhi Wang, Lei Zhang, and Hua Huang. Learning tensor low-rank prior for hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12006–12015, 2021. 2

[52] Xuanyu Zhang, Yongbing Zhang, Ruiqin Xiong, Qilin Sun, and Jian Zhang. Herosnet: Hyperspectral explicable re-

construction and optimal sampling deep network for snapshot compressive imaging. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17511–17520, 2022. 1, 3

[53] Haiyu Zhao, Yuanbiao Gou, Boyun Li, Dezhong Peng, Jiancheng Lv, and Xi Peng. Comprehensive and delicate: An efficient transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14122–14132, 2023. 2

[54] Min Zhao, Longbin Yan, and Jie Chen. Lstm-dnn based autoencoder network for nonlinear hyperspectral image unmixing. *IEEE Journal of Selected Topics in Signal Processing*, 15:295–309, 2021. 1, 2

[55] Yin-Ping Zhao, Jiancheng Zhang, Yongyong Chen, Zhen Wang, and Xuelong Li. Rcump: Residual completion unrolling with mixed priors for snapshot compressive imaging. *IEEE Transactions on Image Processing*, 33:2347–2360, 2024. 7

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