Progressive Content-Aware Coded Hyperspectral Snapshot Compressive Imaging

Xuanyu Zhang, Bin Chen, Wenzhen Zou, Shuai Liu, Yongbing Zhang, *Senior Member, IEEE*, Ruiqin Xiong, *Senior Member, IEEE*, and Jian Zhang, *Member, IEEE*

***Abstract*— Hyperspectral imaging plays a pivotal role across diverse applications, like remote sensing, medicine, and cytology. The utilization of 2D sensors to acquire 3D hyperspectral images (HSIs) via a coded aperture snapshot spectral imaging (CASSI) system has proven successful, owing to its hardware-friendly implementation and fast sampling speed. Nevertheless, for less spectrally sparse scenes, the use of a single snapshot and unreasonable coded aperture design limits the efficacy of CASSI systems and renders HSI reconstruction more ill-posed, leading to compromised spatial and spectral fidelity. This paper proposes a novel Progressive Content-Aware CASSI (PCA-CASSI) framework, which progressively captures HSIs using multiple optimized content-aware coded apertures and fuses all snapshot measurements for reconstruction. By unlocking the full potential of CASSI systems and elevating their performance ceilings, this framework offers researchers new avenues for improving imaging quality. Furthermore, we develop the RndHRNet, a Range-Null space Decomposition (RND)-inspired deep unfolding network with multiple iterative phases for HSI recovery. Each unfolded recovery phase efficiently exploits the physical information within the coded apertures via explicit RND and adaptively explores the spatial-spectral correlation by dual transformer blocks. Through comprehensive experiments, our approach demonstrates superior performance compared to existing state-of-the-art methods in both the multiple- and single-shot compressive HSI imaging tasks with substantial improvements. Code is available at https://github.com/xuanyuzhang21/PCA-CASSI.**

***Index Terms*— Spectral snapshot compressive imaging, deep unfolding, progressive content-aware sampling.**

I. Introduction

HYPERSPECTRAL images (HSIs) embody rich spectral bands and detailed information than conventional RGB images, which have gained growing attention in recent years [1], [2], [3], [4]. Inspired by the compressive sensing (CS) theory [5], [6], [7], the CASSI system aims to utilize a 2D detector to capture 3D hyperspectral scenes. Due to the merits of fast acquisition speed, low sampling cost, and high data throughput, it plays an indispensable role in a wealth of applications, such as remote sensing, object detection, autonomous driving, agricultural inspection, super-resolution, and medical diagnosis [8], [9], [10], [11], [12], [13], [14].

However, in certain specialized applications demanding high imaging accuracy, the information captured by a single snapshot may be insufficient and bring great challenges to HSI reconstruction. For instance, in microscopy and medical imaging, the precise recovery of nucleus speckles and tissue details in hyperspectral imaging is of paramount importance for accurate biological and medical pathology diagnosis [15]. In areas like remote sensing [16], agriculture [17], and water source pollution monitoring [18], the mostly adopted single-shot-based spectral imaging systems may restrict the scope for further improvements in imaging performance, thus posing a challenge for accurately locating the important events and phenomena. Therefore, to guarantee the accuracy of recovered images, capturing the same scene with multiple shots can be necessary and imperative. Remarkably, advancements in sampling devices, such as the enhanced digital micromirror device (DMD) [19] and CCD sensors, enable imaging systems to record multiple snapshot measurements and modify coded patterns with only a marginal increase in acquisition time.

Although previous works like MS-CASSI [20] have explored the potential of multiple shots for snapshot HSI acquisition, there are still several problems that need to be resolved. **Firstly**, the key to enhancing multiple snapshot imaging systems lies in designing optimal multiple-coded apertures. Our focus centers on two essential aspects: **(1)** The coded apertures are expected to exhibit excellent anisotropy and complementarity. Complementary coded apertures facilitate the fusion and interaction of multiple snapshots. Otherwise, they may tend to interfere with each other and undermine the beneficial information; **(2)** Furthermore, the coded apertures are supposed to be adjusted contextually. Prior information from the previous shots enables coded apertures to perceive the HSI content, thus increasing the flexibility and performance of imaging systems. To the best of our knowledge, these two aspects remain under investigation. **Secondly**, existing HSI recovery networks generally focus on single-disperser CASSI system (SD-CASSI). The practical solution for multiple-shot reconstruction is deficient. How to

Manuscript received 8 February 2024; revised 28 April 2024; accepted 27 May 2024. Date of publication 6 June 2024; date of current version 27 November 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62331011 and in part by Shenzhen Research Project under Grant JCYJ20220531093215035. This article was recommended by Associate Editor Z. Xiang. *(Corresponding author: Jian Zhang.)*

Xuanyu Zhang, Bin Chen, and Jian Zhang are with the School of Electronic and Computer Engineering, Peking University, Shenzhen 518055, China (e-mail: xuanyuzhang21@stu.pku.edu.cn; chenbin@stu.pku.edu.cn; zhangjian.sz@pku.edu.cn).

Wenzhen Zou and Yongbing Zhang are with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China (e-mail: 21s151114@stu.hit.edu.cn; ybzhang08@hit.edu.cn).

Shuai Liu is with Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China (e-mail: liu-s20@mails.tsinghua.edu.cn).

Ruiqin Xiong is with the School of Computer Science, Peking University, Beijing 100871, China (e-mail: rqxiong@pku.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2024.3409421.

Digital Object Identifier 10.1109/TCSVT.2024.3409421

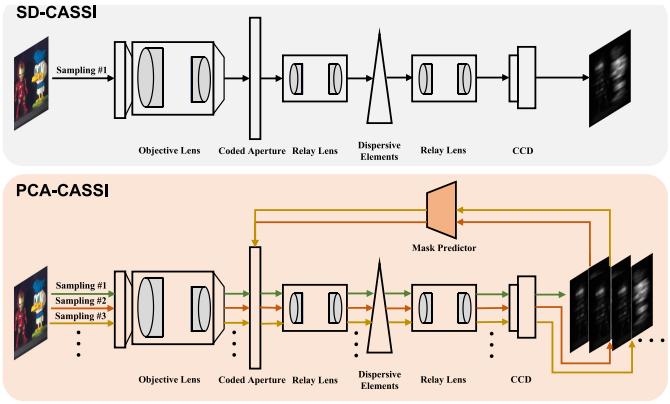


Fig. 1. Illustration of SD-CASSI and the proposed PCA-CASSI system. SD–CASSI compresses the HSI via a single coded aperture, while our PCA-CASSI can progressively capture the same scene with multiple content-aware coded apertures (produced by adaptively learned deep mask predictors), which significantly improves final imaging quality.

utilize the coded apertures to retain the range space information (or measurements) and recover the null space component of HSI has been ignored.

To solve the above-mentioned issues, we propose a novel **P**rogressive **C**ontent-**A**ware **C**oded **A**perture **S**napshot **S**pectral **I**maging (**PCA-CASSI**) framework. It implements content-aware sampling and addresses the requirements for precise multiple-shot spectral imaging (Fig. 1). From a hardware perspective, PCA-CASSI can be easily deployed on SD-CASSI optics. From the perspective of algorithm design, PCA-CASSI is composed of a lightweight recovery network and a small set of deep network-based mask predictors. Furthermore, an **R**ange-**N**ull space **D**ecomposition-inspired **H**yperspectral **R**econstruction **Net**work (**RndHRNet**) is developed to adaptively fuse all the coded measurements. Overall, our contributions can be summarized as follows:

❑ (1) We introduce a novel PCA-CASSI framework, which follows a typical "Encoder+Decoder" paradigm. It captures HSIs through the "Encoder", while the "Decoder" handles the reconstruction process.

❑ (2) A progressive content-aware sampling strategy is proposed to perceive HSI information and optimize the coded aperture for each shot based on measurements from the previous shots.

❑ (3) An $\mathcal{R}-\mathcal{N}$ decomposition-inspired network dubbed RndHRNet is presented for HSI recovery. It uses the Range-Null space Decomposition Module (RNDM) to iteratively refine the null space of HSIs and utilizes the Spectral Spatial Fusion Module (SSFM) to exploit non-local spatial-spectral information.

❑ (4) Experimental results demonstrate the superiority of our method over other state-of-the-art approaches in both multiple- and single-shot HSI reconstruction tasks, with substantial performance improvements.

II. Background

A. Hyperspectral Imaging Systems

Conventional hyperspectral cameras usually make a trade-off between temporal and spectral resolution. Inspired by the compressive sensing (CS) theory, CASSI aims to capture hyperspectral images (HSIs) within a snapshot time. In what follows, we will review several typical works on imaging optical path and coded aperture design.

1) Imaging Optical Path: The most representative CASSI system [21] utilized a single disperser to encode the spatial and spectral information. To improve the imaging performance and information throughput, Kittle et al. [20] captured the same hyperspectral scenes via varied coded apertures. In addition, by temporally aligning the single-shot CASSI system with an RGB camera, Wang et al. [23] provided multi-modal supervision and complementary information for HSI reconstruction. Recently, Lin et al. [24], [25] utilized two high-speed spatial light modulators (SLMs) to realize dual-optical coding. Although the optical path design of CASSI has achieved remarkable results, ensuring a high compression ratio and excellent image quality is still a significant challenge for the community.

2) Coded Aperture Design: Wu et al. [19] first introduced the digital micromirror device (DMD) into the imaging systems to realize flexible and fast programmable coding. Benefiting from the rapid response of DMD, the theory of Restricted Isometry Property (RIP) [26] was introduced to guide the coded aperture optimization. To make the coded aperture retain enough useful information and remove sampling redundancy, Zhang et al. [27] combined the coded aperture optimization and HSI reconstruction with a united network to achieve efficient learning for the mask. However, as the requirement and target scene vary, designing adaptive, content-aware, and data-driven coded apertures becomes worthy of exploration.

B. HSI Reconstruction Methods

1) Model-Based Methods: The traditional model-based methods employ the regularization term inspired by the image prior to solving the ill-posed inverse problem iteratively with widely-used optimization frameworks, such as ISTA [28], [29], GAP [30], etc [31], [32]. Simultaneously, to improve the representation ability of the HSI model, TV [30], GMM [33], [34], image sparsity representation [35] and rank minimization [36], [37] are embedded into the optimization frameworks to provide prior information. Although these methods produce well-recovered results in specific scenarios, it is difficult to design suitable hand-crafted prior terms for all scenes.

2) Deep Learning-Based Methods: Owing to the powerful representation capabilities of deep networks, learning-based HSI reconstruction methods [38] have received increasing attention. To improve the accuracy and perceptual quality, residual blocks [39], spatial-spectral attention modules [40], long-short-term memory units [41], and Fourier domain constraint [42] are embedded into the structure of convolution neural networks (CNNs). Although they can capture the local image patterns, CNNs fail to exploit the global correlation and long-range dependencies. Recently, thanks to the wide application of vision transformers, spatial-wise and channel-wise transformers have been incorporated into the multi-scale encoder-decoder architectures [43]. To alleviate the weak interpretability of existing end-to-end networks,

some researchers have attempted to combine optimization algorithms and deep network priors, such as plug-and-play (PnP) frameworks [44], [45], [46]. Meanwhile, deep unfolding networks [27], [47], [48], [49], [50], [51] have been prevalent for their well-defined architecture. Although these methods have achieved great success, little attention has been paid to multiple-shot HSI reconstruction, limiting the scope of imaging accuracy.

III. Proposed Method

A. Review of the CASSI System

In the CASSI system, the 3D hyperspectral cube is first modulated via a coded aperture and then dispersed via a dispersive element (Fig. 1). Given an HSI sequence $\{\mathbf{X}_i\}_{i=1}^{C} \in \mathbb{R}^{H\times W}$, each frame is modulated via a coded aperture $\mathbf{M} \in \mathbb{R}^{H\times W}$ as:

$$\mathbf{X}'_i = \mathbf{M} \odot \mathbf{X}_i \in \mathbb{R}^{H\times W}, \tag{1}$$

where $\mathbf{X}'_i$ is the i^{th} modulated HSI frame and $\odot$ denotes the Hadamard product. The modulated HSI frames $\{\mathbf{X}'_i\}_{i=1}^{C} \in \mathbb{R}^{H\times W}$ with different wavelengths are then shifted spatially and element-wise summed, leading to a coded measurement as:

$$\mathbf{Y}(m,n) = \sum_{i=1}^{C} \mathbf{X}'_i(m, n+d_i) + \mathbf{N}(m,n), \tag{2}$$

where (m, n) denote the spatial coordinates, and d_i denotes the shifting distance of the i^{th} channel. $\mathbf{N}$ and $\mathbf{Y} \in \mathbb{R}^{H\times(W+C-1)}$ represent the noise and the coded HSI measurement, respectively. Following TSA-Net [40], we assume $\mathbf{N}$ as a shot noise. Specifically, shot noise is influenced by the camera sensor's dynamic range and its quantum efficiency (QE). The measurement affected by shot noise, denoted as $\mathbf{Y}_{sn}$, can be modeled using the equation $\mathbf{Y}_{sn} = \mathcal{B}(\mathbf{Y}/QE, QE)$. Here, $\mathbf{Y}$ represents the noise-free measurement, consisting of integer values ranging from 0 to $2^k - 1$, where k indicates the sensor's bit depth. The function $\mathcal{B}(n, p)$ denotes the binomial distribution, and QE denotes the sensor's quantum efficiency. We set QE and k to 0.4 and 11, respectively. The vectorized form of CASSI can be expressed as:

$$\mathbf{y} = \Phi\mathbf{x} + \mathbf{n} \in \mathbb{R}^p, \tag{3}$$

where $\mathbf{x} \in \mathbb{R}^q$, $\mathbf{y}$, and $\mathbf{n} \in \mathbb{R}^p$ denote the vectorized form of $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{N}$, respectively, with $p = H\times(W + C - 1)$ and $q = H\times W\times C$. Here, $\Phi \in \mathbb{R}^{p\times q}$ represents the sensing matrix.

The classic CASSI system can be generalized to a multiple-shot version. If we capture the same scene $\mathbf{X}$ through N multiple different code apertures $\{\mathbf{M}_i\}_{i=1}^{N}$, every shot can be treated as an implementation of single-shot CASSI system. Hereby, the sampling process is formulated as follows:

$$\left[\mathbf{y}_1^\top, \mathbf{y}_2^\top, \cdots, \mathbf{y}_N^\top\right]^\top = \left[\Phi_1^\top, \Phi_2^\top, \cdots, \Phi_N^\top\right]^\top \mathbf{x} + \mathbf{n}, \tag{4}$$

where $\mathbf{y}_i \in \mathbb{R}^p$ and $\Phi_i = \text{Mask2Mat}(\mathbf{M}_i) \in \mathbb{R}^{p\times q}$ denote the observed HSI measurement and sensing matrix of the i^{th} shot, respectively. Mask2Mat is an operation transforming the physical mask $\mathbf{M}_i$ to its equivalent sensing matrix form Φ_i. The acquisition process of multiple-snapshot compressive imaging remains a linear model and can still be expressed as the simple form of $\mathbf{y} = \Phi\mathbf{x} + \mathbf{n}$.

TABLE I

High-Level Functional Comparison Among Three Representative Imaging Systems [20], [21], [22] for HSI and Our Proposed Progressive Content-Aware PCA-CASSI

Method	SD-CASSI [21]	DCD [22]	MS-CASSI [20]	PCA-CASSI (Ours)
Single CCD sensor	✓	×	✓	✓
Multiple mask switch	×	×	✓	✓
Progressive sampling	×	×	×	✓
Content-aware mask	×	×	×	✓

B. Overview of the Proposed PCA-CASSI

As illustrated in Fig. 2, our PCA-CASSI is a novel "Encoder + Decoder" framework, which aims to compress HSIs by various content-aware learned masks to increase the information throughput and fuse all the captured snapshots for joint HSI reconstruction. Following [52] and [43], this work focuses on HSI sampling and reconstruction, while our contributions may be extensible to other imaging systems like [20], [21], and [22]. For software algorithm, is composed of two sub-modules, namely the mask predictors $\{\mathcal{H}_M^{(i)}\}_{i=2}^{N}$ and reconstruction network $\mathcal{H}_R$. Here, the mask predictors can learn to generate mask contextually in the current shot from the measurement in the previous shot as follows:

$$\mathbf{M}_i = \mathcal{H}_M^{(i)}(\mathbf{y}_{i-1}) \in \mathbb{R}^{H\times W}, \quad i \in \{2, 3, \cdots\}, \tag{5}$$

where $\mathbf{M}_i$ denotes the adaptively predicted mask in i^{th} shot and $\mathbf{y}_{i-1}$ represents the measurement from $(i-1)^{th}$ shot. $\mathbf{M}_1$ is a learnable parameter component. We will elaborate on this module in Sec. III-C. Simultaneously, $\mathcal{H}_R$ is designed to integrate all the coded snapshots based on the $\mathcal{R} - \mathcal{N}$ Decomposition (RND) theory. It is formulated as:

$$\hat{\mathbf{x}} = \mathcal{H}_R(\mathbf{y}_1, \mathbf{M}_1, \cdots, \mathbf{y}_N, \mathbf{M}_N) \in \mathbb{R}^q, \tag{6}$$

where $\hat{\mathbf{x}}$ denotes the reconstructed HSI result. More details are provided in Sec. III-D. For hardware realization, the proposed PCA-CASSI can be easily implemented via the acquisition devices of SD-CASSI [21]. Owing to the rapid response time of 2D sensors and the fast prediction speed of $\mathcal{H}_M^{(i)}$, the increased time caused by multiple shots is indeed limited and can even be negligible for many biological/medical/agricultural applications. Considering its superior performance, we would argue that it is still worth sacrificing a little sampling speed for higher imaging accuracy. As illustrated in Tab. I, our framework is distinguished from other existing imaging systems. Specifically, different from MS-CASSI [20], which uses parallel multiple non-adaptive coded apertures, the proposed PCA-CASSI achieves progressive and sequential sampling, and can also be easily extended to any required number of shots. Different from DCD [22], [53], which uses dual cameras to capture snapshots and grayscale images simultaneously, our framework needs only one camera and is more flexible with respect to the coded apertures. Hence, the proposed PCA-CASSI is hardware-friendly in optics implementation, content-aware in coded aperture design, and effective for HSI imaging.

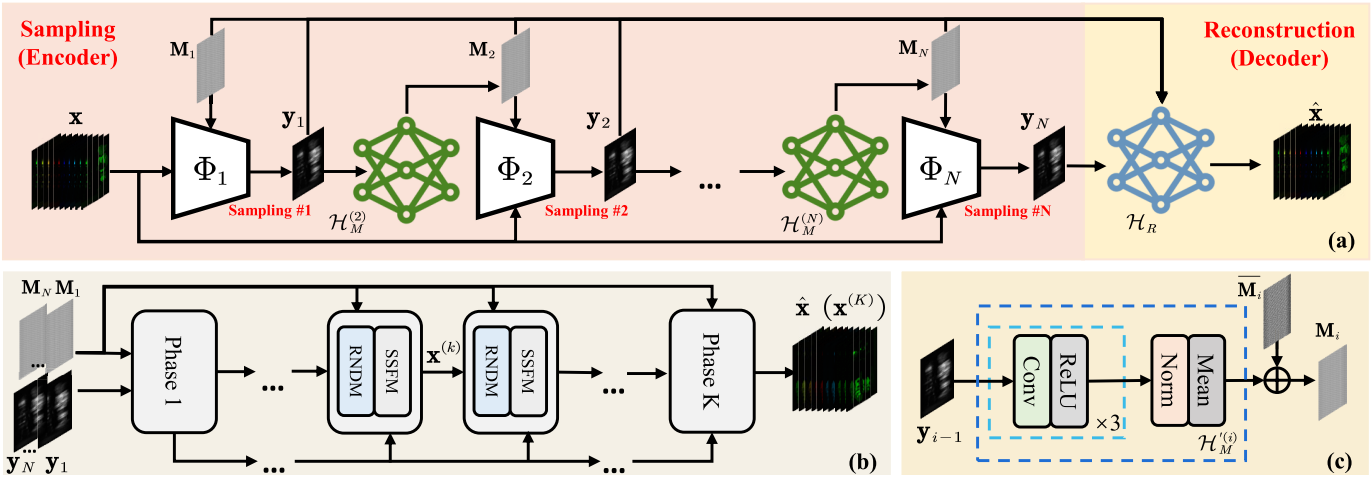


Fig. 2. Illustration of the proposed PCA-CASSI. (a) The overall sampling-reconstruction architecture. (b) The reconstruction process of $\mathcal{H}_R$. (c) The design of mask predictor $\mathcal{H}_M^{(i)}$ for the i^{th} shot. $\mathcal{H}_R$ takes multiple snapshots $\{\mathbf{y}_i\}_{i=1}^N$ and coded apertures $\{\mathbf{M}_i\}_{i=1}^N$ to generate reconstructed HSI $\hat{\mathbf{x}}$, which includes K RND-inspired recovery phases. It is also adaptable to single-shot imaging tasks. $\mathcal{H}_M^{(i)}$ takes $\mathbf{y}_{i-1}$ from previous shot to predict $\mathbf{M}_i$ for current shot.

C. Progressive Content-Aware Sampling

Apart from the fusion mechanism in the restoration process, coded aperture design plays a pivotal role in PCA-CASSI, which enables different measurements to contain complementary information and maintain excellent anisotropy. Considering that different HSIs have various spectral correlations and spatial sparsities, the coded measurement from the previous shot is utilized to optimize the coded aperture in the current shot. Thereby, the optimized coded apertures can perceive the information of HSIs and make content-aware adjustments adaptively. As the shot epoch progresses, the optimized coded apertures tend to be more reasonable with more informative reconstruction results.

Specifically, as shown in Fig. 2 (a), the hyperspectral scene $\mathbf{x}$ is firstly captured by a learned mask $\mathbf{M}_1$ to obtain the measurement $\mathbf{y}_1$. Then, $\mathbf{y}_1$ is adopted to furnish prior information about the target scene and predict subsequent mask $\mathbf{M}_2$ by $\mathcal{H}_M^{(2)}$. Due to the guidance from $\mathbf{y}_1$, the predicted $\mathbf{M}_2$ can perceive HSI content and update the coded pattern. Generally, the coded aperture $\mathbf{M}_i$ in the i^{th} shot is generated from the measurement $\mathbf{y}_{i-1}$ via mask predictor $\mathcal{H}_M^{(i)}$. To be noted, the optimized mask in each shot is required to reflect both the shared properties of the imaging system and the independent characteristics of each HSI. Hence, the masks $\{\mathbf{M}_i\}_{i=1}^N$ are decoupled into two components: a shared component and a content-aware component. The shared components $\{\overline{\mathbf{M}_i}\}_{i=1}^N$ are learnable parameters and jointly optimized with the network. The content-aware component $\overline{\mathbf{M}_i}$ is generated based on the previous measurement $\mathbf{y}_{i-1}$ via a lightweight deep network module $\mathcal{H}_M^{\prime(i)}$, composed of three Conv-ReLU layers, a normalization layer and a mean layer (Fig. 2 (c)). The normalization layer transforms all pixels to the interval [0, 1], while the mean layer calculates channel-wise means of features to produce a mask. This generation pipeline can be formulated as:

$$\begin{aligned}\mathbf{M}_i &= \mathcal{H}_M^{(i)}(\mathbf{y}_{i-1}) \quad \in \mathbb{R}^{H\times W} \\ &= \overline{\mathbf{M}_i} + \eta^{(i)}\mathcal{H}_M^{\prime(i)}(\mathbf{y}_{i-1}),\end{aligned} \tag{7}$$

where $\eta^{(i)}\in\mathbb{R}$ is a learnable scalar for stabilizing training process. Note that $\mathbf{M}_1$ is a learned parameter that is directly learned during the process of network training. It will learn the characteristics of all sampling processes and spectral data, thereby obtaining a coded aperture that is universally applicable and shared across all HSI scenes. Through our progressive content-aware sampling, each snapshot is expected to adaptively learn to capture the complementary HSI information and spectral features.

D. Architecture of RndHRNet

The proposed RndHRNet is devoted to recovering the HSI $\mathbf{x}$ from one or more snapshots $\{\mathbf{y}_i\}_{i=1}^N$ and use coded apertures $\{\mathbf{M}_i\}_{i=1}^N$ to guide the transmission of spectral-spatial features. To start with the noise-free degradation model $\mathbf{y} = \Phi\mathbf{x}$, the HSI reconstruction is formulated as:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - \Phi\mathbf{x}\|_2^2 + \lambda\psi(\mathbf{x}) \in \mathbb{R}^q, \tag{8}$$

where $\mathbf{y} = \left[\mathbf{y}_1^\top, \mathbf{y}_2^\top, \cdots, \mathbf{y}_N^\top\right]^\top \in \mathbb{R}^{pN}$ is the concatenation of all measurement vectors. The first part is the data fidelity term, and the second part $\lambda\psi(\mathbf{x})$ denotes the regularization term generally assumed to be convex. Traditional proximal gradient descent (PGD) algorithm [54] solves Eq. 8 by iterating between the following gradient descent and proximal mapping steps:

$$\mathbf{z}^{(k)} = \mathbf{x}^{(k-1)} - \rho\Phi^\top(\Phi\mathbf{x}^{(k-1)} - \mathbf{y}) \in \mathbb{R}^q, \tag{9}$$

$$\mathbf{x}^{(k)} = \text{prox}_{\lambda\psi}(\mathbf{z}^{(k)}) \in \mathbb{R}^q, \tag{10}$$

where $\mathbf{z}^{(k)}$ and $\mathbf{x}^{(k)}$ denote the intermediate result and reconstruction image in the k^{th} phase; k and ρ denote the number of PGD iterations and the step size, respectively. The limitation of the above Eq. 9 is that gradient descent only finds an approximate sub-optimal solution to the data fidelity term in each iteration, which may not satisfy $\mathbf{y} \equiv \Phi\mathbf{z}^{(k)}$. To overcome the weakness of gradient descent and enhance recovery, inspired by [55] and [56], the $\mathcal{R}-\mathcal{N}$ Decomposition (RND) is introduced to explicitly maintain the consistency constraint

of the data fidelity term. Given a sensing matrix Φ and its pseudo-inverse matrix $\Phi^\dagger = \Phi^\top(\Phi\Phi^\top)^{-1}$, which satisfies that $\Phi\Phi^\top$ is assumed to be invertible [30] and $\Phi\Phi^\dagger = \mathbf{I}$, we have the following theorem for arbitrary HSIs:

Theorem 1 ($\mathcal{R}-\mathcal{N}$ Decomposition [55]): Let $\mathcal{P}_r \triangleq \mathbf{\Phi}^\dagger\mathbf{\Phi}$ be the operator that projects the sample $\mathbf{x}$ from sample domain to the range of $\mathbf{\Phi}^\dagger$, and denote by $\mathcal{P}_n \triangleq \left(\boldsymbol{I} - \mathbf{\Phi}^\dagger\mathbf{\Phi}\right)$ the operator that projects $\mathbf{x}$ to the null space of Φ. Then $\forall\mathbf{x} \in \mathbb{R}^{H\times W\times C}$, there exists the unique decomposition:

$$\mathbf{x} \equiv \mathcal{P}_r(\mathbf{x}) + \mathcal{P}_n(\mathbf{x}) \in \mathbb{R}^q. \tag{11}$$

Here, $\mathcal{P}_r(\mathbf{x})$ and $\mathcal{P}_n(\mathbf{x})$ respectively denote the range and null space of $\mathbf{x}$. The HSI reconstruction can be treated as solving these two components $\mathcal{P}_r(\mathbf{x})$ and $\mathcal{P}_n(\mathbf{x})$. Reconstructing the range space of the HSI signal is to ensure data consistency with respect to measurement $\mathbf{y}$ while refining the null space of the HSI signal aims to remove artifacts and enrich details. Therefore, the core advantage of RND lies in its ability to improve perceptual quality while maintaining reconstruction fidelity. Furthermore, by substituting $\mathbf{y} = \mathbf{\Phi x}$ into Eq. 11, we have the following RND for $\mathbf{x}$:

$$\mathbf{x} \equiv \Phi^\dagger\mathbf{y} + \left(\mathbf{I} - \Phi^\dagger\Phi\right)\mathbf{x} \in \mathbb{R}^q. \tag{12}$$

In HSI reconstruction task, the range-space component $\Phi^\dagger\mathbf{y}$ can actually be calculated from $\mathbf{y}$, while the null-space component $\left[\left(\mathbf{I} - \Phi^\dagger\Phi\right)\mathbf{x}\right]$ can be refined and estimated by the network. To be noted, the combination of $\Phi^\dagger\mathbf{y}$ and any solution $\mathbf{s}$ for the null-space component projected by $\left(\mathbf{I} - \Phi^\dagger\Phi\right)$ strictly enjoys the exact data consistency, *i.e.* $\forall\mathbf{s},\ \Phi\left[\Phi^\dagger\mathbf{y} + \left(\mathbf{I} - \Phi^\dagger\Phi\right)\mathbf{s}\right] \equiv \mathbf{y}$. Thus, our motivation is to estimate the null-space component $\left[\left(\mathbf{I} - \Phi^\dagger\Phi\right)\mathbf{x}\right]$ only and remain the clean range-space component unchanged to alleviate the recovery difficulty. Furthermore, we implement the null-space learning in Eq. 12 by a deep network and resort to the proximal mapping to refine the null-space component iteratively as follows:

$$\begin{aligned}\mathbf{z}^{(k)} &= \Phi^\dagger\mathbf{y} + \left(\mathbf{I} - \Phi^\dagger\Phi\right)\mathbf{x}^{(k-1)} \\ &= \mathbf{x}^{(k-1)} + \Phi^\dagger(\mathbf{y} - \Phi\mathbf{x}^{(k-1)}) \in \mathbb{R}^q, \end{aligned}\tag{13}$$

$$\mathbf{x}^{(k)} = \arg\min_{\mathbf{x}} \frac{1}{2}\left\|\mathbf{x} - \mathbf{z}^{(k)}\right\|_2^2 + \lambda\psi(\mathbf{x}) \in \mathbb{R}^q, \tag{14}$$

where $\mathbf{z}^{(k)}$ and $\mathbf{x}^{(k)}$ respectively denote the result of RND-based update and proximal mapping in the k^{th} iteration. Inspired by Eqs. 13 and 14, in the following subsections, we develop a range-null decomposition module (RNDM) and spatial-spectral fusion module (SSFM) for HSI recovery.

1) Range-Null Space Decomposition Module: Based on Eq. 13, our RNDM aims to retain the range space and refine the null space of the HSIs iteratively (Fig. 3). Given snapshots $\{\mathbf{y}_i\}_{i=1}^N$ and masks $\{\mathbf{M}_i\}_{i=1}^N$, RNDM can yield the intermediate result $\mathbf{z}^{(k)} \in \mathbb{R}^q$ adaptively as:

$$\mathbf{z}^{(k)} = \mathrm{RNDM}^{(k)}(\mathbf{x}^{(k-1)}, \mathbf{y}_1, \cdots, \mathbf{y}_N, \mathbf{M}_1, \cdots, \mathbf{M}_N). \tag{15}$$

Due to the cumulative errors caused by equipment deviation, noise corruption, and alignment of the continuous spectrum in real scenarios, using physical masks $\{\mathbf{M}_i\}_{i=1}^N$ directly may raise the risk of physics imprecision. Therefore, a deep mask enhanced module $\mathcal{H}_{EM}^{(k)}$ is introduced to produce two enhanced deep mask representations $\mathbf{F}_i^{(k)} \in \mathbb{R}^{H\times W\times C}$ and $\mathbf{E}_i^{(k)} \in \mathbb{R}^{H\times W\times C}$ from the physical mask $\mathbf{M}_i \in \mathbb{R}^{H\times W}$, which corrects the bias between the degradation process and the physical mask via an attention mechanism [43] as follows:

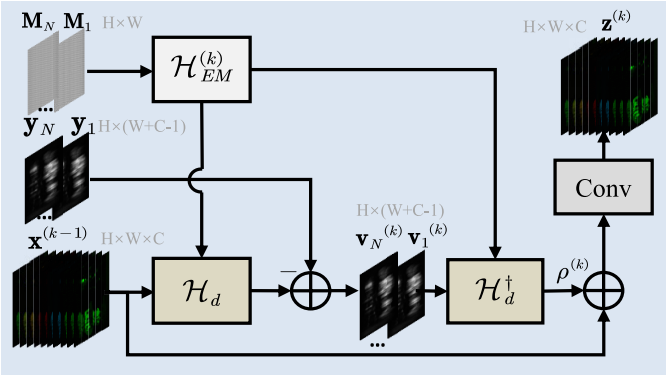


Fig. 3. Details of our RNDM. It retains the range-space component and recovers the null-space component of HSIs.

$$\mathbf{F}_i^{(k)}, \mathbf{E}_i^{(k)} = \mathcal{H}_{EM}^{(k)}(\mathbf{M}_i). \tag{16}$$

Furthermore, the crucial step of RNDM is to solve the degradation operator $\mathcal{H}_d$ and its pseudo-inverse operator $\mathcal{H}_d^\dagger$. Specifically, $\mathcal{H}_d$ simulates the process of mask modulation, dispersion, and compression in Sec. III-A with the current state $\mathbf{x}^{(k-1)}$ and enhanced mask representations $\mathbf{F}_i^{(k)}$. $\mathcal{H}_d^\dagger$ is designed to provide an initialization from 2D signals to 3D cubes with the enhanced representation $\mathbf{E}_i^{(k)}$. To dynamically balance the contribution of range- and null-space signals, the learnable parameter $\rho^{(k)}$ is incorporated into the optimization process. Hence, we can do $\mathcal{R}-\mathcal{N}$ decomposition on each $\mathbf{y}_i$ to get the intermediate results $\{\mathbf{z}_i^{(k)}\}_{i=1}^N$. Then, a 1×1 convolution is utilized to adaptively merge $\{\mathbf{z}_i^{(k)}\}_{i=1}^N$ into $\mathbf{z}^{(k)}$ as follows, which is then fed into the subsequent proximal mapping module.

$$\begin{aligned}\mathbf{v}_i^{(k)} &= \mathbf{y}_i - \mathcal{H}_d(\mathbf{x}_i^{(k-1)}, \mathbf{F}_i^{(k)}) \in \mathbb{R}^p, \\ \mathbf{z}_i^{(k)} &= \mathbf{x}_i^{(k-1)} + \rho^{(k)}\mathcal{H}_d^\dagger(\mathbf{v}_i^{(k)}, \mathbf{E}_i^{(k)}) \in \mathbb{R}^q, \\ \mathbf{z}^{(k)} &= \mathrm{Conv}^{(k)}(\left[\mathbf{z}_1^{(k)}, \cdots, \mathbf{z}_N^{(k)}\right]) \in \mathbb{R}^q, \end{aligned}\tag{17}$$

where $\mathbf{v}_i^{(k)}$ and $[\cdot]$ denote an auxiliary variable and the channel-wise concatenation, respectively. The measurements compressed by various coded patterns can reflect different HSI contents and have complementary information.

2) Spectral-Spatial Fusion Module: To implement Eq. 14 with deep networks, SSFM is adopted to refine the fused intermediate result $\mathbf{z}^{(k)}$ and yield the HSI reconstruction result $\mathbf{x}^{(k)}$ (Fig. 4). To explicitly make our SSFM able to explore the spectral and spatial priors in a high-throughput manner, we designed a multi-scale U-shaped proximal subnetwork, which consists of a Spectral-Spatial Fusion Block (SSFB), a Feature Interaction Module (FIM), a downsampling and upsampling operation, and a fusion operation. The spectral-spatial fusion block (SSFB) plays a crucial role in extracting spectral correlation and local spatial

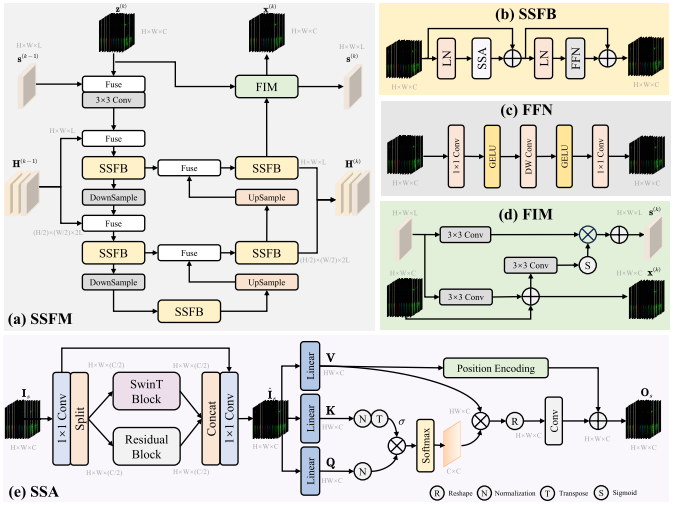


Fig. 4. Details of the proposed SSFM. It extracts multiscale spectral-spatial features with high-throughput transmission. (a) The overall architecture of our Spectral-Spatial Fusion Module (SSFM). (b) Spectral-Spatial Fusion Block (SSFB). (c) Feed-Forward Network (FFN). (d) Feature Interaction Module (FIM). (e) Spectral-Spatial self-Attention block (SSA).

information, which includes two Layer Normalization (LN) layers, a Spectral-Spatial self-Attention block (SSA), and a Feed-Forward Network (FFN). Given that HSI representations are spatially sparse and spectrally correlated, capturing spatial interactions and spectral correlation are just as important. Inspired by the previous works [43], [57], the channel-wise and spatial-wise transformer blocks are plugged into the U-shaped architecture. As illustrated in Fig. 4 (e), the spatial-wise transformer block aims to introduce local biases and spatial cues for the long-range transformer blocks via the sliding-window-based self-attention mechanism and residual blocks. For convenience, we assume that the batch size and head number are set to 1, respectively. Specifically, the input feature of size $H \times W \times C$ is firstly processed via an 1×1 convolution and split in the channel dimension as follows:

$$\mathbf{I}_s^{(1)}, \mathbf{I}_s^{(2)} = \text{Split}(\text{Conv}(\mathbf{I}_s)). \tag{18}$$

One part of the split feature is fed to the Swin transformer block [58], [59] to explore the spatial correlation and non-local information. Additionally, the other part of the feature is fed to the residual block to capture local image cues. Finally, the two parts of features are fused adaptively via a 1×1 convolution as follows:

$$\hat{\mathbf{I}}_s = \text{Conv}\left(\left[\text{SwinT}(\mathbf{I}_s^{(1)}), \text{RB}(\mathbf{I}_s^{(2)})\right]\right) + \mathbf{I}_s, \tag{19}$$

where SwinT and RB denote the sliding-window-based transformer and residual block. Then, the spectral-wise transformer encodes each processed feature frame into a token and explores spectral self-attention. Concretely, the spectral-wise transformer block encodes each feature frame $\hat{\mathbf{I}}_s \in \mathbb{R}^{H\times W\times C}$ to the channel-wise tokens via three linear layers to obtain the value $\mathbf{V} \in \mathbb{R}^{HW\times C}$, key $\mathbf{K} \in \mathbb{R}^{HW\times C}$ and query $\mathbf{Q} \in \mathbb{R}^{HW\times C}$:

$$\mathbf{Q} = \hat{\mathbf{I}}_s \mathbf{W}_\mathbf{Q}, \mathbf{K} = \hat{\mathbf{I}}_s \mathbf{W}_\mathbf{K}, \mathbf{V} = \hat{\mathbf{I}}_s \mathbf{W}_\mathbf{V}, \tag{20}$$

where $\mathbf{W}_\mathbf{Q}$, $\mathbf{W}_\mathbf{K}$ and $\mathbf{W}_\mathbf{V} \in \mathbb{R}^{C\times C}$ respectively denote the learnable parameters. Then, $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V} \in \mathbb{R}^{HW\times C}$ are split into N heads along the channel dimension, namely $\mathbf{Q} = [\mathbf{Q}_1, \cdots, \mathbf{Q}_N]$, $\mathbf{K} = [\mathbf{K}_1, \cdots, \mathbf{K}_N]$ and $\mathbf{V} = [\mathbf{V}_1, \cdots, \mathbf{V}_N]$. Note that the channel dimension of each head is C/N. Then, we calculate the self-attention map $\mathbf{A}_j \in \mathbb{R}^{H\times W\times (C/N)}$ for each head j as follows:

$$\mathbf{A}_j = \mathbf{V}_j \,\text{Softmax}(\sigma_j \mathbf{K}_j^\top \mathbf{Q}_j), \quad j \in \{1, 2, \cdots, N\}, \tag{21}$$

where $\sigma_j \in \mathbb{R}$ is a learnable parameter to represent the variation of spectral intensity. Finally, the output is reshaped and fused via a convolution and added with position encoding $\mathcal{H}_{pos}$.

$$\mathbf{O}_s = [\mathbf{A}_1, \cdots, \mathbf{A}_N]\,\mathbf{W} + \mathcal{H}_{pos}(\mathbf{V}) \in \mathbb{R}^{H\times W\times C}, \tag{22}$$

where $\mathbf{W} \in \mathbb{R}^{C\times C}$ is the learnable weights of convolution. $\mathcal{H}_{pos}$ is composed of several depth-wise convolutions and GELU activation.

To avoid information loss and model degradation, a feature interaction module (FIM) [27] is incorporated into the SSFM to interact with the spatial-spectral features in other phases, which will further extract the hidden spectral feature $\mathbf{s}^{(k)} \in \mathbb{R}^{H\times W\times L}$ and produce the reconstructed HSI result $\mathbf{x}^{(k)} \in \mathbb{R}^{H\times W\times C}$ through several 3×3 convolutions and a pixel-wise attention mechanism. Furthermore, since the feature maps at each scale of the U-shaped architecture also have well-preserved spatial information, the multi-scale features $\mathbf{H}^{(k)} = [\mathbf{h}_0^{(k)}, \mathbf{h}_1^{(k)}, \cdots, \mathbf{h}_{D-1}^{(k)}]$ produced by SSFB in all layers are also utilized for feature fusion, where $D = 2$ and $L = 32$ respectively represent the number of layers in the U-shaped network and the channel number of spectral features, $\mathbf{h}_i^{(k)} \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times 2^i L}$ denotes the intermediate results produced by SSFB in the i^{th} layer of the k^{th} phase. The "Fuse" operation is implemented via the concatenation operation and a 1×1 convolution. The cascaded features $\mathbf{H}^{(k-1)}$ and a hidden spectral feature $\mathbf{s}^{(k-1)} \in \mathbb{R}^{H\times W\times L}$ in the previous phase are empirically validated to be conducive for the reconstruction of the current phase. Formally, our SSFM can be formulated as:

$$\mathbf{H}^{(k)}, \mathbf{s}^{(k)}, \mathbf{x}^{(k)} = \text{SSFM}^{(k)}(\mathbf{z}^{(k)}, \mathbf{s}^{(k-1)}, \mathbf{H}^{(k-1)}). \tag{23}$$

E. Network Slimming Strategy

Furthermore, we propose a network slimming strategy to achieve a more lightweight and parameter-efficient reconstruction process. Specifically tailored to deep unfolding networks, our flexible network slimming strategy is to retain the main components of each recovery phase in the RndHR-Net as shared across all K phases while maintaining the interaction components between different phases that are weight-independent, such as FIM. We have empirically validated that this strategy can significantly reduce the parameter number of our network without compromising its performance, even with an increased number of phases. Moreover, these weight-independent modules remarkably enhance the capacity of network and representation capabilities compared to a globally shared structure.

F. Network Training and Implementation

Without bells and whistles, our recovery network $\mathcal{H}_R$ and mask predictors $\{\mathcal{H}_M^{(i)}\}_{i=2}^N$ can be jointly optimized end-to-end.

Specifically, given training data $\{\mathbf{x}_i\}_{i=1}^{N_d}$, our loss function is defined as:

$$\mathcal{L}(\Theta) = \frac{1}{HWCN_d}\sum_{i=1}^{N_d}\|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_2^2, \tag{24}$$

where $\hat{\mathbf{x}}_i$ and N_d are the result and sample number. $\Theta = \left[\bigcup_{i\in\{2,\cdots,N\}}\mathcal{H}_M^{(i)}\right]\cup\mathcal{H}_R$ denotes the set of all parameters with $\mathcal{H}_M^{(i)} = \{\overline{\mathbf{M}}_i, \eta^{(i)}, \mathcal{H}_M'^{(i)}\}$, $\mathcal{H}_R = \{\text{RNDM}^{(k)}, \text{SSFM}^{(k)}\}_{k=1}^{K}$, and $\text{RNDM}^{(k)} = \{\rho^{(k)}, \mathcal{H}_{EM}^{(k)}, \text{Conv}^{(k)}\}$. The Adam optimizer is employed for network training of 200 epochs. The learning rate is initialized to 4×10^{-4} and scheduled to 1×10^{-6} using cosine annealing.

G. Relationships to Other Works

The proposed PCA-CASSI represents a significant advancement over prior methods. Unlike MS-CASSI [20], which utilizes multiple random fixed binary masks with a parallel sampling mechanism, our approach incorporates sequentially progressive sampling and employs innovative, learnable masks that dynamically adjust based on image content and current reconstruction outputs. In addition, while MS-CASSI relies on conventional Total Variation (TV) for reconstructing hyperspectral images (HSIs), PCA-CASSI leverages advanced RND-based deep unfolding networks, enhancing the accuracy of HSI reconstruction. In contrast to HerosNet [27], which employs a single optimized coded aperture for single-shot sampling, PCA-CASSI is designed to generate varying coded apertures for multiple samplings. These apertures are dynamic and adapt in response to content, offering a substantial improvement in versatility and efficiency. Furthermore, while both MST [43] and our method utilize mask-guided attention, PCA-CASSI extends this concept to multiple samplings and integrates it with Range-Null space decomposition to achieve a more refined mask representation. This not only aligns more closely with the physical principles of sampling but also differs fundamentally from MST, which uses the attention map as merely a directional guide for the model to focus on areas yielding high-fidelity HSI representations. Our approach also surpasses MST by incorporating not only spectral-wise self-attention but also a novel spatial self-attention mechanism, implemented through a sliding window. This enhancement enables our network to more effectively discern spatial-spectral correlation. Compared to Deep RND [55], PCA-CASSI introduces a mutable mask representation grounded in range-null space decomposition, offering greater flexibility and improved efficiency for spectral SCI tasks. Given that Deep RND has not been previously applied to spectral SCI, we have developed specialized operators, $\mathcal{H}_d^{\dagger}$ and $\mathcal{H}_d$, specifically tailored to emulate the physical imaging processes inherent in spectral SCI. In summary, PCA-CASSI significantly expands upon existing designs in optical path configuration, mask optimization, and deep unfolding reconstruction algorithms, setting a new standard for content-aware spectral SCI.

TABLE II
SETTINGS OF OUR DIFFERENT RNDHRNET VARIANTS

Network Variant	Progressive Content-Aware Sampling	Weight-Sharing
RndHRNet-M	×	✓
RndHRNet-L	×	×
RndHRNet-PM	✓	✓
RndHRNet-PL	✓	×

IV. EXPERIMENTAL RESULTS

A. Experimental Settings

In this paper, the effectiveness of the proposed method has been validated on both simulated and real datasets. Following [40], simulation experiments are conducted on the public HSI datasets CAVE [62] and KAIST [63] with size $256\times256\times28$ (*i.e.*, $C = 28$). Meanwhile, the five real HSI measurements [51] with size 660×714 are used for test. The metrics of PSNR and SSIM [64] are employed to evaluate the final HSI reconstruction quality quantitatively. We present the settings of different variants of the proposed RndHRNet in Tab. II. Throughout this paper, "-M" and "-L" denote the mobile and large versions of RndHRNet. "-P" denotes the RndHRNet with our proposed progressive content-aware sampling scheme. "-KPh" denotes the deep unfolding network with K phases.

B. Comparison on Simulated Data

To verify and demonstrate the effectiveness of proposed PCA-CASSI and RndHRNet, we compare our methods with other state-of-the-art (SOTA) methods, including a model-based methods [30], six CNN-based methods [39], [40], [41], [42], [51], [61], and five recent transformer networks [43], [50], [52], [60], [65].

1) Single-Shot Reconstruction: Following the settings of previous works [40], [51], we conduct the single-shot reconstruction on the KAIST dataset [63]. As reported in Tabs. III and V, the proposed RndHRNet-M and RndHRNet-L with 10 phases yields 38.93 and 38.54 dB on PSNR, which significantly outperforms most of the competing SOTA approaches [43], [52], [60], [61] **over 2 dB**. Meanwhile, we also observe that in relatively lightweight settings such as 2 phases, our RndHRNet-M-2Ph can still outperform the two most recent SOTA methods DAUHST-2Ph [50] by 0.7 dB and RDLUF-2Ph [65] by 0.5 dB. Remarkably, when applying our network slimming strategy, RndHRNet-M surpasses the weight-independent RndHRNet-L and **achieves a significant PSNR score of 37.04 dB with a total of 0.98 M parameters**. As depicted in Fig. 5, our approach excels in recovering sharper edges and more realistic details, attesting to the efficacy of our proposed RNDM and SSFM in preserving data consistency and effectively exploiting non-local correlations.

2) Multiple-Shot Reconstruction: We extend RndHRNet to a multiple-shot imaging setting with our proposed progressive content-aware sampling strategy. For RndHRNet-PL/-PM, two optimized and content-aware masks are dynamically generated and utilized to capture HSIs progressively. For other SOTA methods, we employ two fixed real masks $\mathbf{M}$ and $(\mathbf{1}-\mathbf{M})$ [40] to realize two-shot imaging. To adapt existing end-to-end

TABLE III

PSNR (dB)/SSIM COMPARISON RESULTS OF THE PROPOSED RNDHRNETS AND ELEVEN STATE-OF-THE-ART HSI RECONSTRUCTION METHODS ON SINGLE-SHOT IMAGING TASK. THE BEST AND SECOND-BEST SCORES ARE HIGHLIGHTED IN **BOLD** AND UNDERLINED, RESPECTIVELY

Test set	Scene01	Scene02	Scene03	Scene04	Scene05	Scene06	Scene07	Scene08	Scene09	Scene10	Average
GAP-TV [30]	26.82	22.89	26.31	30.65	23.64	21.85	23.76	21.98	22.63	23.10	24.36
(ICIP 2016)	0.739	0.665	0.802	0.852	0.703	0.663	0.688	0.655	0.682	0.584	0.669
DeSCI [36]	27.13	23.04	26.62	34.96	23.94	22.38	24.45	22.03	24.56	23.59	25.27
(TPMAI 2018)	0.748	0.620	0.818	0.897	0.706	0.683	0.743	0.673	0.732	0.587	0.721
TSA-Net [40]	32.03	31.00	32.25	39.19	29.39	31.44	30.32	29.35	30.01	29.59	31.46
(ECCV 2020)	0.892	0.858	0.915	0.953	0.884	0.908	0.878	0.888	0.890	0.874	0.894
DGSMP [51]	33.26	32.09	33.06	40.54	28.86	33.08	30.74	31.55	31.66	31.44	32.63
(CVPR 2021)	0.915	0.898	0.925	0.964	0.882	0.937	0.886	0.923	0.911	0.925	0.917
SRN [39]	34.85	35.11	35.89	42.12	32.53	34.59	33.52	32.63	35.04	31.99	34.82
(ArXiv 2021)	0.937	0.935	0.949	0.975	0.944	0.955	0.924	0.947	0.944	0.938	0.945
HDNet [42]	35.14	35.67	36.03	42.30	32.69	34.46	33.67	32.48	34.89	32.38	34.97
(CVPR 2022)	0.935	0.940	0.943	0.969	0.946	0.952	0.926	0.941	0.942	0.937	0.943
MST [43]	35.40	35.87	36.51	42.27	32.77	34.80	33.66	32.67	35.39	32.50	35.18
(CVPR 2022)	0.941	0.944	0.953	0.973	0.947	0.955	0.925	0.948	0.949	0.941	0.948
MST++ [60]	35.80	36.24	37.39	43.85	33.41	35.43	34.35	33.71	36.67	33.38	36.02
(CVPRW 2022)	0.943	0.947	0.957	0.973	0.952	0.957	0.934	0.953	0.953	0.945	0.951
CST [52]	35.16	35.60	36.57	42.29	32.82	35.15	33.85	33.52	35.28	32.84	35.31
(ECCV 2022)	0.938	0.942	0.953	0.972	0.948	0.956	0.927	0.952	0.946	0.940	0.947
GAP-Net [61]	33.74	33.26	34.28	41.03	31.44	32.40	32.27	30.46	33.51	30.24	33.26
(IJCV 2023)	0.911	0.900	0.929	0.967	0.919	0.925	0.902	0.905	0.915	0.895	0.917
BIRNAT [41]	36.79	37.89	40.61	**46.94**	35.42	35.30	36.58	33.96	39.47	32.80	37.58
(TPAMI 2023)	0.951	0.957	0.971	0.985	0.964	0.959	0.955	0.956	0.970	0.938	0.960
RndHRNet-M	36.57	37.74	39.29	44.76	34.56	35.97	35.58	34.13	38.21	33.61	37.04
(2 Phases)	0.951	0.959	0.965	0.979	0.960	0.963	0.948	0.958	0.963	0.947	0.959
RndHRNet-M	**37.81**	**40.40**	**42.09**	45.88	**37.34**	**37.42**	**37.97**	35.08	**40.61**	**34.74**	**38.93**
(10 Phases)	**0.963**	**0.975**	**0.975**	**0.985**	**0.976**	**0.973**	**0.965**	0.967	**0.975**	**0.960**	**0.971**
RndHRNet-L	36.15	37.34	38.47	43.95	34.57	35.82	35.37	33.95	37.57	33.46	36.66
(2 Phases)	0.948	0.956	0.963	0.975	0.959	0.961	0.943	0.957	0.961	0.945	0.957
RndHRNet-L	37.29	40.07	41.48	45.59	36.08	37.34	37.27	**35.55**	39.99	34.69	38.54
(10 Phases)	0.959	0.974	0.972	0.985	0.971	0.972	0.960	**0.970**	0.972	0.960	0.969

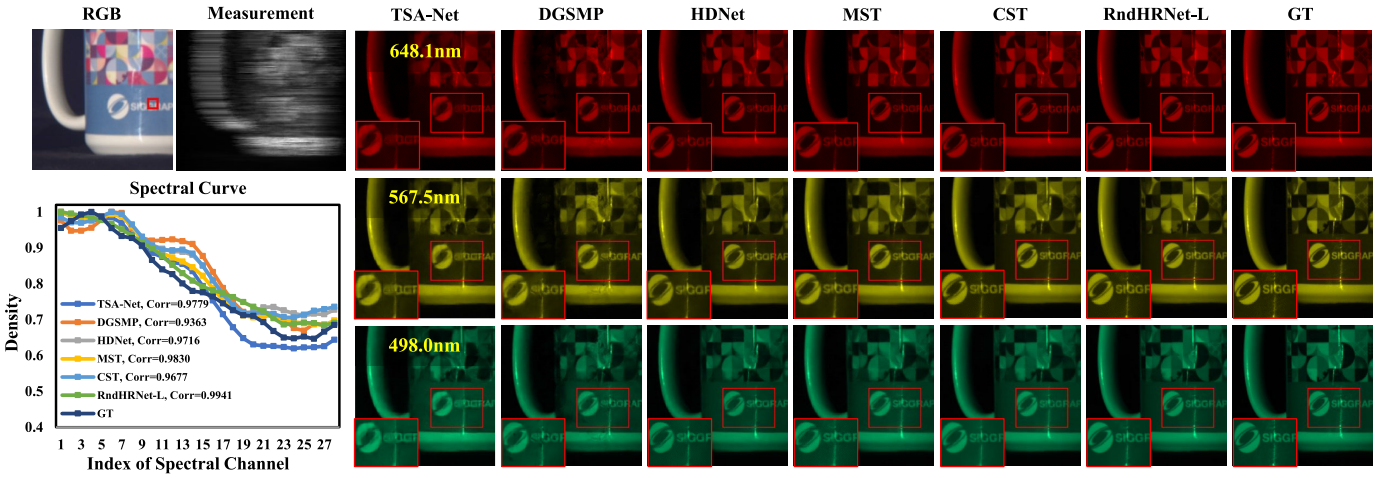


Fig. 5. Visual comparison of our RndHRNet-L and other state-of-the-art (SOTA) methods in single-shot imaging case.

deep learning-based methods to multiple-shot reconstruction, we utilize a few 1×1 convolutions to fuse two measurements and masks in learning-based methods (SRN, HDNet, MST, and CST). Besides, for model-based methods (GAP-TV and ADMM-TV), we jointly solve two sub-problems corresponding to two compressed measurements. To adapt other deep unfolding networks like DAUHST [50] and RDLUF [65] to the two-shot setting, we use linear projection or RDLGD block to process $\{\mathbf{y}_1, \Phi_1\}$ and $\{\mathbf{y}_2, \Phi_2\}$ respectively, and fuse the intermediate results using a 1×1 convolution.

Tab. IV shows the improvements achieved by all reconstruction methods compared to the single-shot imaging results listed in Tab. III. However, we observe that certain methods do not fully exploit the complementary information of multiple masks, thus falling short of reaching their full performance potential. Thanks to our RND-inspired fusion mechanism, RndHRNet-L with 2 phases achieves 38.39 dB on PSNR and

TABLE IV

PSNR (dB)/SSIM Comparison Results of the Proposed RndHRNets and Eleven State-of-the-Art HSI Reconstruction Methods on Two-Shot Imaging Task. The Best and Second-Best Scores Are Highlighted in **Bold** and Underlined, Respectively

Test set	Scene01	Scene02	Scene03	Scene04	Scene05	Scene06	Scene07	Scene08	Scene09	Scene10	Average
GAP-TV [30]	27.62	25.92	23.65	34.20	23.90	23.97	23.46	23.68	25.18	24.22	25.58
(ICIP 2016)	0.739	0.665	0.762	0.872	0.708	0.670	0.682	0.654	0.708	0.589	0.705
ADMM-TV [30]	27.85	25.65	23.84	33.56	23.94	23.85	23.60	23.93	25.04	24.54	25.58
(ICIP 2016)	0.768	0.684	0.791	0.886	0.732	0.698	0.712	0.695	0.743	0.603	0.731
TSA-Net [40]	34.79	35.09	34.50	40.94	32.21	35.63	32.06	31.37	32.27	30.42	33.06
(ECCV 2020)	0.930	0.920	0.930	0.955	0.925	0.938	0.902	0.919	0.913	0.887	0.917
SRN [39]	34.79	35.09	34.50	40.94	32.21	35.63	33.63	34.45	34.27	33.82	34.93
(ArXiv 2021)	0.930	0.920	0.930	0.955	0.925	0.938	0.919	0.933	0.930	0.942	0.932
HDNet [42]	35.20	36.00	35.01	41.19	32.76	36.10	33.44	34.96	34.42	33.95	35.30
(CVPR 2022)	0.941	0.945	0.936	0.968	0.946	0.958	0.917	0.958	0.937	0.958	0.947
MST [43]	35.50	36.00	36.38	43.16	33.31	35.91	33.70	35.07	36.15	34.50	35.97
(CVPR 2022)	0.942	0.945	0.943	0.975	0.946	0.954	0.916	0.951	0.944	0.954	0.946
MST++ [60]	35.89	36.70	36.44	43.56	33.18	36.52	33.96	35.58	35.89	34.67	36.24
(CVPRW 2022)	0.947	0.946	0.947	0.975	0.948	0.960	0.918	0.961	0.945	0.958	0.951
CST [52]	35.57	35.99	35.94	42.32	33.25	36.36	34.02	35.62	35.52	34.83	35.94
(ECCV 2022)	0.939	0.940	0.943	0.965	0.945	0.955	0.921	0.954	0.942	0.956	0.946
RndHRNet-M	37.45	39.64	38.83	44.84	35.71	38.24	35.94	37.68	38.63	36.44	38.34
(2 Phases)	0.960	0.970	0.959	0.983	0.965	0.971	0.944	0.973	0.965	0.970	0.966
RndHRNet-L	37.49	39.78	39.01	44.49	35.62	38.47	36.27	37.58	38.69	36.54	38.39
(2 Phases)	0.961	0.969	0.957	0.980	0.965	0.970	0.945	0.971	0.963	0.971	0.965
RndHRNet-PM	38.39	40.90	40.88	48.07	36.70	39.24	37.27	38.82	40.42	37.60	39.83
(2 Phases)	0.964	0.976	0.962	0.989	0.970	0.977	0.953	0.977	0.969	0.974	0.971
RndHRNet-PL	**39.49**	**43.52**	**40.91**	**47.44**	**38.45**	**40.15**	**38.72**	**39.48**	**40.35**	**38.78**	**40.73**
(2 Phases)	**0.970**	**0.984**	**0.968**	**0.988**	**0.978**	**0.980**	**0.963**	**0.981**	**0.973**	**0.981**	**0.977**

TABLE V

Comparisons of Our Proposed Method With Two Most Recent SOTA Approaches in the One-Shot Setting

Method	DAUHST [50]	RDLUF [65]	RndHRNet-M	RndHRNet-L
Phase	2	2	2	2
#Params.(M)	1.40	1.21	**0.98**	1.82
PSNR(dB)	36.34	36.54	**37.04**	36.66
Phase	9	9	10	10
#Params(M).	6.15	**1.85**	2.97	9.48
PSNR(dB)	38.36	**39.57**	38.93	38.54

TABLE VI

Comparisons of Our Proposed Method With Two Most Recent SOTA Approaches in the Two-Shot Setting

Method	DAUHST [50]	RDLUF [65]	RndHRNet-PM	RndHRNet-PL
Phase	2	2	2	2
#Params.(M)	1.41	1.22	**0.99**	1.84
PSNR(dB)	35.64	37.19	39.83	**40.73**

0.965 on SSIM, surpassing the single-shot case by 1.73 dB. It is worth noting that even when utilizing the exact same system and masks, our proposed RndHRNet-L outperforms the SOTA methods RDLUF [65] and MST++ [60] by 1.20 dB and 2.15 dB on PSNR, respectively. Considering the trade-off between computational complexity and performance, RndHRNet with 10 phases is not conducted on the multiple-shot case. Furthermore, through the use of content-aware optimized masks, our proposed PCA-CASSI achieves impressive results, reaching 40.73 dB on PSNR and 0.977 on SSIM. This is a significant improvement compared to two SOTA end-to-end learning-based methods CST [52] and MST [43], where our method outperforms them by 4.79 dB and 4.76 dB on PSNR, respectively, validating the effectiveness of our imaging framework. We also compared our method with two recent SOTA deep unfolding methods, DAUHST [50] and RDLUF [65], as reported in Tab. VI. Our method outperforms DAUHST by 5.09 dB in the two-shot setting. Additionally, despite RDLUF's integration of advanced transformer designs, our RndHRNet-L still surpasses RDLUF by 3.54 dB in the two-shot scenario. This is credited to our powerful progressive content-aware sampling scheme, facilitating scene-adaptive mask learning, along with the specially designed RndHRNet for two-shot reconstruction. Furthermore, as depicted in Fig. 6, the HSIs produced by our methods exhibit clearer spatial details and more accurate spectral consistency.

3) Visualization of Learned Masks: Fig. 7 shows our learned masks and measurements captured for different shots. An interesting finding is that the optimized mask $\mathbf{M}_2$ contains a shared pattern $\overline{\mathbf{M}}_2$ while containing content-aware textures $\mathcal{H}_M'^{(2)}(\mathbf{y}_1)$, thus facilitating the capture of a more comprehensive range of spectral and spatial information. Moreover, upon analyzing the Fourier spectra $\mathcal{F}(\Phi_1^\dagger \mathbf{y}_1)$ and $\mathcal{F}(\Phi_2^\dagger \mathbf{y}_2)$, we note that the first shot prioritizes the acquisition of low-frequency information and the overall content of HSIs. Conversely, the second shot emphasizes high-frequency cues and intricate textures. This empirical evidence confirms the distinct focus of different shots on various components of HSIs, validating the efficacy of our progressive information acquisition through the sampling mechanism.

4) Computational Complexity: As shown in Tab. VII and Fig. 8, RndHRNet-L achieves better PSNR scores than that

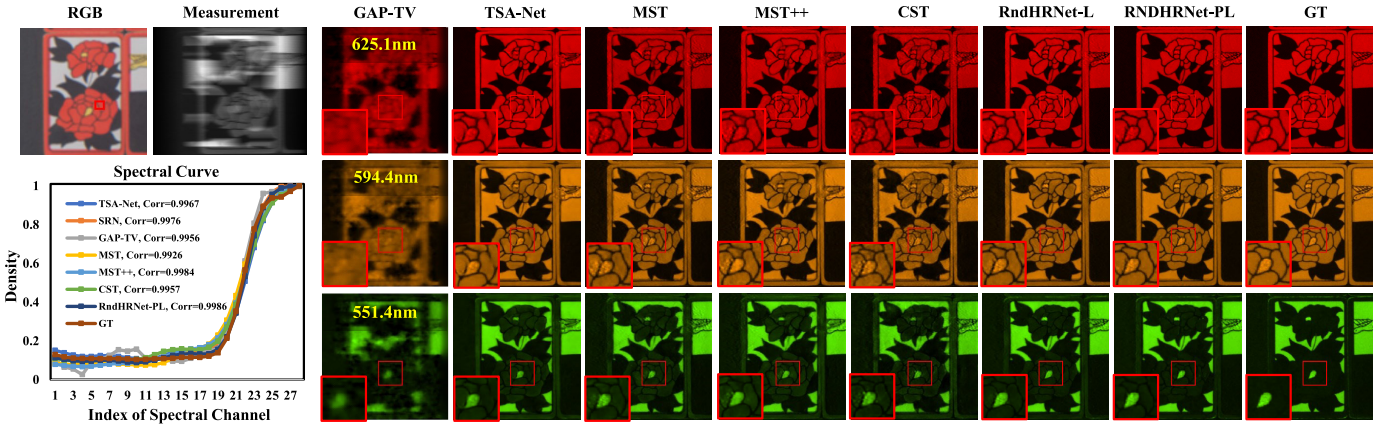


Fig. 6. Visual comparison of our proposed method and other SOTA methods on multiple-shot HSI reconstruction task.

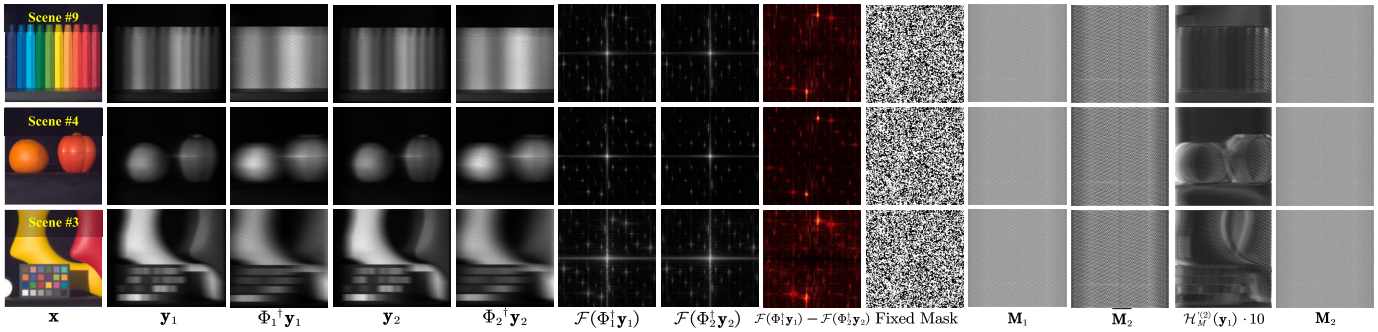


Fig. 7. Visualization results of measurements $\mathbf{y}_1$, $\mathbf{y}_2 \in \mathbb{R}^p$ with $p = H \times (W + C - 1)$ captured in different shots and $\mathbf{M}_1$, $\mathbf{M}_2 \in \mathbb{R}^{H\times W}$, where $\mathbf{M}_2 = \overline{\mathbf{M}_2} + \eta^{(2)} \cdot \mathcal{H}_M'^{(2)}(\mathbf{y}_1)$. Compared to the fixed mask used in prior studies, our learnable masks are more compact and evenly distributed, and reflect the spectral-spatial features of different HSI scenes. We also present the mean of Fourier spectrum of $\Phi_1^\dagger \mathbf{y}_1$, $\Phi_2^\dagger \mathbf{y}_2 \in \mathbb{R}^q$ with $q = H \times W \times C$ in the spectral dimension, along with their corresponding residual maps, where $\mathcal{F}$ denotes the operation of Fast Fourier Transform (FFT), followed by the absolute value computation. For clearer presentation, all the element values of residual $\mathcal{H}_M'^{(2)}(\mathbf{y}_1) \in \mathbb{R}^{H\times W}$ are scaled by a factor of 10.

TABLE VII

EVALUATION OF THE COMPUTATIONAL COMPLEXITY AND PERFORMANCE ON MULTIPLE-/SINGLE-SHOT RECONSTRUCTION TASKS. HERE, "-*K*PH" DENOTES THAT THE NUMBER OF DEEP UNFOLDED ITERATIVE PHASES IN THE NETWORK IS SET TO *K*

Method	Shot	Test Time (ms)	#Params. (M)	PSNR (dB)	SSIM
HDNet [42]	1	116.92	2.37	34.97	0.943
MST [43]	1	176.36	2.03	35.18	0.948
CST [52]	1	153.18	1.36	35.31	0.947
RndHRNet-M-2Ph	1	130.42	**0.98**	37.04	0.959
RndHRNet-L-2Ph	1	189.13	1.82	36.66	0.957
HDNet [42]	2	117.45	2.37	35.30	0.947
MST [43]	2	179.25	2.03	35.97	0.946
CST [52]	2	159.21	1.36	35.94	0.946
RndHRNet-M-2Ph	2	148.28	**0.99**	38.34	0.966
RndHRNet-PM-2Ph	2	158.12	1.01	39.83	0.971
RndHRNet-L-2Ph	2	199.21	1.82	38.39	0.965
RndHRNet-PL-2Ph	2	249.45	1.84	40.73	0.977

of CST and MST with close inference speeds and even fewer parameters. With scene-adaptive mask predictor, the proposed method with 2 phases can obtain 2.34 dB gains on PSNR with only a 0.02 M increase in parameters, proving the practicality of the progressive content-aware sampling. Moreover, with our proposed network slimming strategy, the parameter number of RndHRNet-M can be further reduced **less than 1M** with even notable improvements in reconstruction

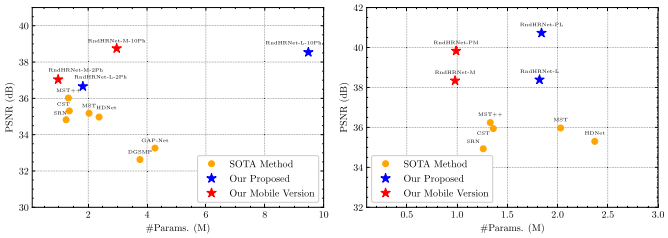


Fig. 8. Comparison of SOTA method and the proposed RndHRNet on parameters and PSNR. The single-shot reconstruction results are reported in the left column, and the two-shot results are presented in the right column.

performance (0.33 dB PSNR gain) in the single-shot setting. Even utilizing 10 phases, the parameter number of RndHRNet can be decreased **from 9.48M to 2.97M**, accompanied by a notable 0.39dB increase on PSNR. We attribute this to the fact that excessively deep networks and overwhelming parameter numbers are not conducive to the learning and optimization of RndHRNet. Similar conclusions hold in the two-shot scenario. Our mobile RndHRNet version achieves outstanding reconstruction performance (39.83 dB on PSNR) with few parameters, providing the opportunity for mobile deployments.

C. Reconstruction From Real Data

To conduct an objective assessment of the proposed RndHRNet-L using real data, we provide comprehensive visualization results with various comparison methods [38], [40],

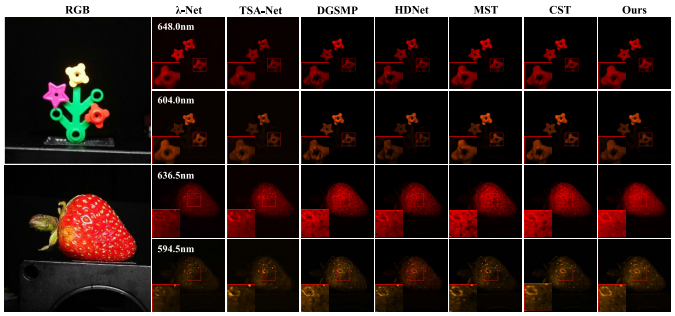


Fig. 9. Real data results of the proposed RndHRNet-L and other SOTA methods on single-shot reconstruction.

TABLE VIII
EXPERIMENT RESULTS OF THE EXTENSION OF OUR PROPOSED PCA-CASSI TO TWO DEEP HSI RECONSTRUCTION ALGORITHMS

Method	#Params.(M)	PSNR (dB)	SSIM
MST	2.03	35.97	0.946
MST (PCA)	2.05 (0.02 ↑)	38.21 (2.24 ↑)	0.964 (0.018 ↑)
CST	1.36	35.94	0.946
CST (PCA)	1.39 (0.03 ↑)	36.09 (0.15 ↑)	0.957 (0.011 ↑)

[42], [43], [51], [52]. To simulate real-world imaging scenarios, we inject 11-bit shot noise into the training data [40]. As depicted in Fig. 9, our recovered results exhibit enhanced perceptual quality, featuring clearer HSI contents and reduced artifacts. Notably, in cropped regions containing flowers, our results display smoother textures, and in cropped regions containing strawberries, there is notably less noise, demonstrating the generalization capability and robustness of our approach. Clearly, our method excels in recovering high-quality HSIs across diverse content and degradation scenes.

D. Extension of Progressive Content-Aware Sampling

To further validate the effectiveness and extensibility of our proposed Progressive Content-Aware (PCA) sampling scheme, we extend this mechanism to two existing SOTA deep reconstruction methods: MST [43] and CST [52]. Tab. VIII demonstrates remarkable improvements brought by PCA introduction. By employing PCA sampling, MST achieves a substantial 2.24 dB gain in PSNR for the two-shot scenario, with only 0.02 M additional parameters. Similarly, CST exhibits gains of 0.15 dB and 0.011 in PSNR and SSIM, respectively. These results affirm the distinctive adaptability of our progressive content-aware sampling, which is capable of enhancing various recovery networks effectively.

E. Ablation Study

1) Ablation Study on the Proposed Sampling Mechanism: To evaluate the contribution of the proposed PCA-CASSI, we deploy the imaging system on different mask combinations and report the results in Tab. IX. For convenience, all the deep reconstruction networks are set as a single-phase RndHRNet.

a) Multiple-shot vs. single-shot: Comparing case 1 and case 4, or case 2 and case 5, HSI reconstruction with two fixed or optimized masks surpasses one-shot reconstruction by 0.69 dB or 2.86 dB on PSNR, suggesting that multiple complementary coded apertures capture richer information.

TABLE IX
EVALUATION OF DIFFERENT MASK COMBINATIONS. HERE, THE RECONSTRUCTION NETWORK INCLUDES 1 PHASE FOR SIMPLICITY

Case	Fixed Mask	Optimized Mask	Content-Aware	PSNR (dB)	SSIM
1	1	0	×	35.04	0.948
2	0	1	×	36.33	0.957
3	1	1	×	38.29	0.967
4	2	0	×	35.73	0.953
5	0	2	×	39.19	0.972
6	0	2	✓	39.62	0.974

TABLE X
EVALUATION OF THE EFFECTIVENESS OF FOUR DIFFERENT COMPONENTS IN OUR DEVELOPED RNDHRNET-L-2PH VARIANTS

Case	RNDM	CTB	STB	FIM	PSNR	SSIM
(a)	×	✓	✓	✓	35.13	0.951
(b)	✓	×	✓	✓	36.15	0.950
(c)	✓	✓	×	✓	36.10	0.953
(d)	✓	✓	✓	×	36.43	0.953
(e)	✓	✓	✓	✓	36.66	0.957

b) Learned vs. fixed masks: As shown in Tab. IX, the more adaptively optimized masks are used, the better the reconstruction performance is, indicating that more learned masks can remove redundancy and retain useful information (case 2,3,5).

c) Content-aware vs. shared: Comparing case 5 and case 6, we find that although the jointly optimized masks are conducive to information acquisition, the same coded apertures are shared and non-adaptive for all HSI scenes, which can lead to serious loss of spectral anisotropy. With the proposed progressive content-aware strategy, HSI reconstruction with dynamically generated content-aware masks can surpass the results with shared optimized masks by 0.43 dB, which proves the effect of our mask optimization algorithm.

2) Ablation Study on the Proposed RndHRNet: To assess the individual contributions of different components in our proposed RndHRNet, we conduct an ablation study on single-shot reconstruction using a two-phase reconstruction network. The focus of this study is on four key modules: the range-nullspace decomposition module (RNDM), the channel-wise transformer block (CTB), the spatial-wise transformer block (STB), and the feature interaction module (FIM). Tab. X presents the PSNR (dB) and SSIM results for various settings. When we remove RNDM and retrain a variant of the proposed model with only two proximal mapping modules, the PSNR experiences a decline of 1.43 dB, affirming the effectiveness of RNDM in the reconstruction process. Additionally, we create two variant models, case (b) and case (c), by respectively removing CTB and STB from SSFM. Without CTB, the PSNR and SSIM values drop by 0.51 dB and 0.007, respectively, while without STB, the results show a decrease of 0.56 dB in PSNR and 0.004 in SSIM. This demonstrates the necessity and significance of cooperative integration of spectral and spatial correlations in the SSFM for high-quality reconstruction. Furthermore, the absence of the proposed FIM results in a reduction of 0.23 dB in PSNR and 0.004 in SSIM, validating the importance of phase interaction for ensuring high throughput in the network. Overall, these

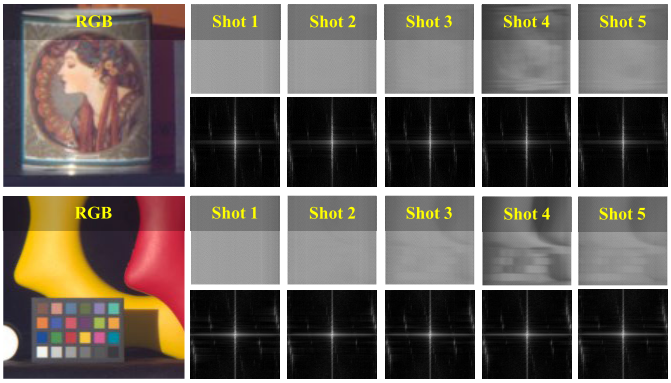


Fig. 10. Visualization of our five learned masks $\{\mathbf{M}_i\}_{i=1}^{N}$ (upper) and the Fourier spectrums of $\{\Phi_i^{\dagger}\mathbf{y}_i\}_{i=1}^{N}$ (lower) regarding two different HSIs from the KAIST dataset.

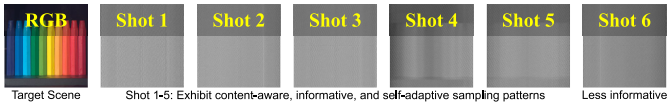


Fig. 11. Visualization of our six learned masks regarding the HSI named "Scene 2" from the KAIST dataset.

TABLE XI

ABLATION STUDIES ON THE EFFECTIVENESS OF OUR DEVELOPED RANGE-NULLSPACE DECOMPOSITION-INSPIRED MODULE

Method	GD	Learned GD	RNDM (Ours)
PSNR (dB)	35.23	35.62	36.66
SSIM	0.948	0.950	0.957

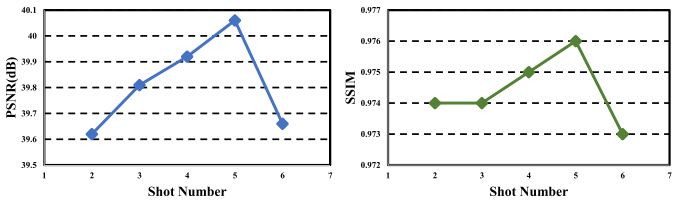


Fig. 12. PSNR (dB) and SSIM evaluation results of our RNDHRNet-PL under settings of different shot numbers.

findings highlight the essential role played by each module in enhancing the reconstruction performance of our RndHRNet.

Meanwhile, we conduct a comparative analysis between our RND-inspired and traditional gradient descent steps. The introduced RND (Eq. 13) explicitly enforces measurement consistency $\Phi\hat{\mathbf{x}} \equiv \mathbf{y}$. Unlike the transpose $\Phi^{\top}$, our provided implementation for the pseudo-inverse $\Phi^{\dagger}$ can offer a non-trivial and more accurate initialization from $\mathbf{y}$, enhancing the mask representation. Ablation studies presented in Tab. XI demonstrate that both the traditional gradient descent (GD) and the incorporation of enhanced mask representation into the gradient descent module (Learned GD) yield inferior performance compared to our proposed RND module. The explicit enforcement of measurement consistency in RND proves to be a critical factor in achieving superior results.

3) Ablation Study on Progressive N-Shot Reconstruction: To explore the performance upper bound of the proposed progressive sampling in multiple-shot reconstruction, we conducted retraining of PCA-CASSI with varying shot numbers, specifically setting N to 2, 3, 4, 5, and 6. The results in terms of PSNR and SSIM are depicted in Fig. 12. As the shot number increases, the reconstruction performances of PCA-CASSI improve correspondingly, peaking when the shot number is set to 5. However, as the shot number increases beyond 5, the PSNR and SSIM results exhibit a slight decline, attributed to the increased network parameters and potential overfitting. Fig. 10 reveals that with an increasing shot number, more HSI contents become reflected in the coded apertures, positively influencing HSI reconstruction. However, an excessive number of progressive samples may pose challenges in network optimization and lead to performance saturation. Moreover, an abundance of samples can contribute to network overfitting. In Fig. 11, for instance, when the shot number $N \geq 6$, the optimized masks fail to capture new HSI contents or provide useful information due to excessive sampling. Further investigations and explanations of this phenomenon will be pursued in our future work.

V. CONCLUSION

In this paper, we present PCA-CASSI, a novel deep learning-based framework for spectral snapshot compressive imaging. Our approach progressively compresses hyperspectral images (HSIs) using content-aware optimized coded apertures and then fuses the snapshots for accurate reconstruction. Inspired by the $\mathcal{R} - \mathcal{N}$ decomposition, we introduce RndHRNet, a deep unfolding network tailored for precise HSI reconstruction. To enhance its representation capabilities, we propose a range-null space decomposition module that iteratively refines the null-space component of HSIs. Additionally, we offer a mobile version of RndHRNet, significantly reducing memory complexity and improving parameter efficiency without compromising reconstruction performance, thereby enabling seamless integration with mobile imaging devices. Extensive experiments demonstrate the superiority and significant improvement potential of our method over other state-of-the-art (SOTA) techniques in both the multiple-/single-shot HSI imaging tasks.

REFERENCES

[1] X. Yuan, D. J. Brady, and A. K. Katsaggelos, "Snapshot compressive imaging: Theory, algorithms, and applications," *IEEE Signal Process. Mag.*, vol. 38, no. 2, pp. 65–88, Mar. 2021.

[2] W. He, N. Yokoya, and X. Yuan, "Fast hyperspectral image recovery of dual-camera compressive hyperspectral imaging via non-iterative subspace-based fusion," *IEEE Trans. Image Process.*, vol. 30, pp. 7170–7183, 2021.

[3] S. Zhang, H. Huang, and Y. Fu, "Fast parallel implementation of dual-camera compressive hyperspectral imaging system," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 11, pp. 3404–3414, Nov. 2019.

[4] H. Chen et al., "Spectral-wise implicit neural representation for hyperspectral image reconstruction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 5, pp. 3714–3727, May 2024.

[5] B. Chen and J. Zhang, "Content-aware scalable deep compressed sensing," *IEEE Trans. Image Process.*, vol. 31, pp. 5412–5426, 2022.

[6] J. Zhang, B. Chen, R. Xiong, and Y. Zhang, "Physics-inspired compressive sensing: Beyond deep unrolling," *IEEE Signal Process. Mag.*, vol. 40, no. 1, pp. 58–72, Jan. 2023.

[7] Y. Hu, Y. Wang, and J. Zhang, "DEAR-GAN: Degradation-aware face restoration with GAN prior," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4603–4615, Sep. 2023.

[8] J. Lei, W. Xie, J. Yang, Y. Li, and C.-I. Chang, "Spectral–spatial feature extraction for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 8131–8143, Oct. 2019.

[9] W. Xie, T. Jiang, Y. Li, X. Jia, and J. Lei, "Structure tensor and guided filtering-based algorithm for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4218–4230, Jul. 2019.

[10] M. Ding, X. Fu, T.-Z. Huang, J. Wang, and X.-L. Zhao, "Hyperspectral super-resolution via interpretable block-term tensor modeling," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 3, pp. 641–656, Apr. 2021.

[11] J. Lei, X. Li, B. Peng, L. Fang, N. Ling, and Q. Huang, "Deep spatial–spectral subspace clustering for hyperspectral image," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2686–2697, Jul. 2021.

[12] L. Wang, Z. Xiong, G. Shi, W. Zeng, and F. Wu, "Simultaneous depth and spectral imaging with a cross-modal stereo system," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 3, pp. 812–817, Mar. 2018.

[13] X. Wang, J. Chen, Q. Wei, and C. Richard, "Hyperspectral image super-resolution via deep prior regularization with parameter estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 1708–1723, Apr. 2022.

[14] X. Yang, W. Cao, Y. Lu, and Y. Zhou, "QTN: Quaternion transformer network for hyperspectral image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 12, pp. 7370–7384, Dec. 2023.

[15] A. Bullen, "Microscopic imaging techniques for drug discovery," *Nature Rev. Drug Discovery*, vol. 7, no. 1, pp. 54–67, Jan. 2008.

[16] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 2, pp. 6–36, Jun. 2013.

[17] B. Lu, P. Dao, J. Liu, Y. He, and J. Shang, "Recent advances of hyperspectral imaging technology and applications in agriculture," *Remote Sens.*, vol. 12, no. 16, p. 2659, Aug. 2020.

[18] M. B. Stuart, A. J. S. McGonigle, and J. R. Willmott, "Hyperspectral imaging in environmental monitoring: A review of recent developments and technological advances in compact field deployable systems," *Sensors*, vol. 19, no. 14, p. 3071, Jul. 2019.

[19] Y. Wu, I. O. Mirza, G. R. Arce, and D. W. Prather, "Development of a digital-micromirror-device-based multishot snapshot spectral imaging system," *Opt. Lett.*, vol. 36, no. 14, pp. 2692–2694, Jul. 2011.

[20] D. Kittle, K. Choi, A. Wagadarikar, and D. J. Brady, "Multiframe image estimation for coded aperture snapshot spectral imagers," *Appl. Opt.*, vol. 49, no. 36, pp. 6824–6833, 2010.

[21] A. Wagadarikar, R. John, R. Willett, and D. Brady, "Single disperser design for coded aperture snapshot spectral imaging," *Appl. Opt.*, vol. 47, no. 10, p. B44, 2008.

[22] T. Zhang, Y. Fu, L. Wang, and H. Huang, "Hyperspectral image reconstruction using deep external and internal learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8558–8567.

[23] L. Wang, Z. Xiong, D. Gao, G. Shi, and F. Wu, "Dual-camera design for coded aperture snapshot spectral imaging," *Appl. Opt.*, vol. 54, no. 4, pp. 848–858, 2015.

[24] X. Lin, G. Wetzstein, Y. Liu, and Q. Dai, "Dual-coded compressive hyperspectral imaging," *Opt. Lett.*, vol. 39, no. 7, pp. 2044–2047, Apr. 2014.

[25] X. Lin, Y. Liu, J. Wu, and Q. Dai, "Spatial–spectral encoded compressive hyperspectral imaging," *ACM Trans. Graph.*, vol. 33, no. 6, pp. 1–11, Nov. 2014.

[26] Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[27] X. Zhang, Y. Zhang, R. Xiong, Q. Sun, and J. Zhang, "HerosNet: Hyperspectral explicable reconstruction and optimal sampling deep network for snapshot compressive imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17511–17520.

[28] J. M. Bioucas-Dias and M. A. T. Figueiredo, "A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2992–3004, Dec. 2007.

[29] J. Zhang, C. Zhao, D. Zhao, and W. Gao, "Image compressive sensing recovery using adaptively learned sparsifying basis via L0 minimization," *Signal Process.*, vol. 103, pp. 114–126, Oct. 2014.

[30] X. Yuan, "Generalized alternating projection based total variation minimization for compressive sensing," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 2539–2543.

[31] J. Ma, X.-Y. Liu, Z. Shou, and X. Yuan, "Deep tensor ADMM-net for snapshot compressive imaging," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10222–10231.

[32] J. Tan, Y. Ma, H. Rueda, D. Baron, and G. R. Arce, "Compressive hyperspectral imaging via approximate message passing," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 2, pp. 389–401, Mar. 2016.

[33] J. Yang et al., "Video compressive sensing using Gaussian mixture models," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4863–4878, Nov. 2014.

[34] J. Yang et al., "Compressive sensing by learning a Gaussian mixture model from measurements," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 106–119, Jan. 2015.

[35] L. Wang, Z. Xiong, G. Shi, F. Wu, and W. Zeng, "Adaptive nonlocal sparse representation for dual-camera compressive hyperspectral imaging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 10, pp. 2104–2111, Oct. 2017.

[36] Y. Liu, X. Yuan, J. Suo, D. J. Brady, and Q. Dai, "Rank minimization for snapshot compressive imaging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 2990–3006, Dec. 2019.

[37] Y. Xu, Z. Wu, J. Chanussot, and Z. Wei, "Hyperspectral computational imaging via collaborative Tucker3 tensor decomposition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 98–111, Jan. 2021.

[38] X. Miao, X. Yuan, Y. Pu, and V. Athitsos, "Lambda-Net: Reconstruct hyperspectral images from a snapshot measurement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4059–4069.

[39] J. Wang, Y. Zhang, X. Yuan, Y. Fu, and Z. Tao, "A simple and efficient reconstruction backbone for snapshot compressive imaging," 2021, *arXiv:2108.07739*.

[40] Z. Meng, J. Ma, and X. Yuan, "End-to-end low cost compressive spectral imaging with spatial–spectral self-attention," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 187–204.

[41] Z. Cheng et al., "Recurrent neural networks for snapshot compressive imaging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2264–2281, Feb. 2023.

[42] X. Hu et al., "HDNet: high-resolution dual-domain learning for spectral compressive imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17521–17530.

[43] Y. Cai et al., "Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17481–17490.

[44] X. Yuan, Y. Liu, J. Suo, and Q. Dai, "Plug-and-play algorithms for large-scale snapshot compressive imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1444–1454.

[45] S. Zheng et al., "Deep plug-and-play priors for spectral snapshot compressive imaging," *Photon. Res.*, vol. 9, no. 2, p. B18, 2021.

[46] Z. Meng, Z. Yu, K. Xu, and X. Yuan, "Self-supervised neural networks for spectral snapshot compressive imaging," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2602–2611.

[47] L. Wang, C. Sun, Y. Fu, M. H. Kim, and H. Huang, "Hyperspectral image reconstruction using a deep spatial–spectral prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8024–8033.

[48] L. Wang, C. Sun, M. Zhang, Y. Fu, and H. Huang, "DNU: Deep non-local unrolling for computational spectral imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1658–1668.

[49] S. Zhang, L. Wang, L. Zhang, and H. Huang, "Learning tensor low-rank prior for hyperspectral image reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12001–12010.

[50] Y. Cai et al., "Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 37749–37761.

[51] T. Huang, W. Dong, X. Yuan, J. Wu, and G. Shi, "Deep Gaussian scale mixture prior for spectral compressive imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16211–16220.

[52] Y. Cai et al., "Coarse-to-fine sparse transformer for hyperspectral image reconstruction," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 686–704.

[53] Y. Fu, T. Zhang, L. Wang, and H. Huang, "Coded hyperspectral image reconstruction using deep external and internal learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3404–3420, Jul. 2022.

[54] J. Zhang and B. Ghanem, "ISTA-net: Interpretable optimization-inspired deep network for image compressive sensing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1828–1837.

[55] D. Chen and M. E. Davies, "Deep decomposition learning for inverse imaging problems," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 510–526.

[56] D. Chen, J. Tachella, and M. E. Davies, "Equivariant imaging: Learning beyond the range space," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Sep. 2021, pp. 4379–4388.

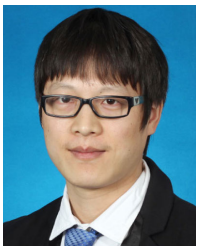
[57] K. Zhang et al., "Practical blind image denoising via Swin-Conv-UNet and data synthesis," 2022, *arXiv:2203.13278*.
[58] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.
[59] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[60] Y. Cai et al., "MST++: Multi-stage spectral-wise transformer for efficient spectral reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 744–754.
[61] Z. Meng, X. Yuan, and S. Jalali, "Deep unfolding for snapshot compressive imaging," *Int. J. Comput. Vis.*, vol. 131, no. 11, pp. 2933–2958, Nov. 2023.
[62] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, "Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2241–2253, Sep. 2010.
[63] I. Choi, D. S. Jeon, G. Nam, D. Gutierrez, and M. H. Kim, "High-quality hyperspectral reconstruction using a spectral prior," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 1–13, Dec. 2017.
[64] W. Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, pp. 600–612, 2004.
[65] Y. Dong, D. Gao, T. Qiu, Y. Li, M. Yang, and G. Shi, "Residual degradation learning unfolding framework with mixing priors across spectral and spatial for compressive spectral imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22262–22271.

Xuanyu Zhang received the B.E. degree from the School of Electrical and Information Engineering, Tianjin University, Tianjin, China. He is currently pursuing the master's degree in computer applications technology with the School of Electronic and Computer Engineering, Peking University, Shenzhen, China. His research interests include spectral compressive imaging and image restoration.

Bin Chen received the B.S. degree from the School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China, in 2021. He is currently pursuing the Ph.D. degree in computer applications technology with the School of Electronic and Computer Engineering, Peking University, Shenzhen, China. His research interests include compressive sensing, image restoration, and computer vision.

Wenzhen Zou received the B.E. degree from the School of Computer Science and Technology, Harbin Institute of Technology, Weihai, China. He is currently pursuing the master's degree in computer technology with Harbin Institute of Technology (Shenzhen), Shenzhen, China. His primary research interests include compressive sensing and computer vision.

Shuai Liu received the B.S. degree from the School of Physical Science and Technology, Inner Mongolia University, China, in 2017, and the M.S. degree from the School of Physics, Beijing Institute of Technology, China, in 2020. He is currently pursuing the Ph.D. degree in control science and engineering with Tsinghua University, China. His research interests include compressive sensing, image restoration, and computational imaging.



Yongbing Zhang (Senior Member, IEEE) received the Ph.D. degree in computer science from Harbin Institute of Technology, Harbin, China, in 2010. He is currently a Professor of computer science and technology with Harbin Institute of Technology (Shenzhen), Shenzhen, China. His research interests include computational imaging, especially exploring new image acquisition and intelligent processing methods and equipment through the deep intersection of machine learning, signal processing, and Fourier optics.

Ruiqin Xiong (Senior Member, IEEE) received the B.S. degree in computer science from the University of Science and Technology of China, Hefei, China, in 2001, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2007. From 2002 to 2007, he was a Research Intern with Microsoft Research Asia. From 2007 to 2009, he was a Senior Research Associate with the University of New South Wales, Sydney, NSW, Australia. In 2010, he joined the School of Electronic Engineering and Computer Science, Peking University, Beijing, where he is currently a Professor. He has authored or coauthored more than 140 technical papers in refereed international journals and conferences. His research interests include image and video processing, statistical image modeling, deep learning, neuromorphic cameras, and computational imaging. He was a recipient of the Best Student Paper Award from the SPIE Conference on Visual Communications and Image Processing in 2005 and the Best Paper Award from the IEEE Visual Communications and Image Processing in 2011. He was a co-recipient of the Best Student Paper Award from the IEEE Visual Communications and Image Processing in 2017.

Jian Zhang (Member, IEEE) received the Ph.D. degree from the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, in 2014. He is currently an Assistant Professor and leads the Visual-Information Intelligent Learning LAB (VILLA), School of Electronic and Computer Engineering, Peking University, Shenzhen, China. He has published over 100 technical papers in refereed international journals and conference proceedings. His research interests include intelligent multimedia processing, including low-level vision, computational imaging, AI-generated content (AIGC), and security. He received the Best Paper Award from the 2011 IEEE Visual Communications and Image Processing (VCIP) and was a co-recipient of the Best Paper Award from the 2018 IEEE MultiMedia. He serves as an Associate Editor for the *Journal of Visual Communication and Image Representation*.