Compact Self-adaptive Coding for Spectral Compressive Sensing

Zhan Shi, Hao Ye, Tao Lv, Yibo Wang, and Xun Cao, *Member, IEEE*

2023 IEEE International Conference on Computational Photography (ICCP) | 979-8-3503-1676-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCP56744.2023.10233711

Abstract—Spectral snapshot compressive imaging (SCI) has been extensively studied and applied to various fields. Although the typical coded aperture snapshot spectral imaging (CASSI) presents an effective paradigm, its fixed code designs do not sufficiently exploit flexible and optimal modulation in consideration of scene sparsity. In this paper, we present a novel **C**ompact **S**elf-adaptive optical **C**oding framework for **S**pectral **C**ompressive **S**ensing, termed **3CS**, to optimize the coded pattern adaptively for better hyper-spectral videos perception. Our framework enables extracting context high-frequency components from the compressed domain without requiring hybrid guiding camera. The specifically designed mask distribution enables higher light efficiency, and is robust against temporal correlation reduction when processing dynamic spectral videos. Extensive experiments and model discussions validate the superiority of the proposed framework over traditional end-to-end (E2E) methods in various aspects for the spectral reconstruction.

Index Terms—Computational Photography, Spectral Compressive Imaging, Adaptive Mask

1 Introduction

Multi/Hyper-spectral cameras capture the spatial distribution of electromagnetic radiation within specific wavelength ranges under the multiple effects of illumination and reflection. It holds tremendous potential for advancing cutting-edge research and applications in fields such as remote sensing, medical imaging, artificial intelligence [1], [2], [3], etc. For dynamic scenes requiring high imaging speed, spectral snapshot compressive imaging (SCI) captures the hyperspectral image (HSI) of the target scene one shot by multiplexing it onto a 2D detector in a decomposable pattern [4], [5]. With recent developments in compressed sensing (CS) theory [6], this problem has become more manageable through use of coded aperture snapshot spectral imaging (CASSI) [7], [8], [9], and HSI reconstruction techniques. For sparse HSI $x \in \mathbb{R}^N$, CASSI implements CS forward process "$y = \Phi x$" using a combination of coding apertures and dispersion elements, where $\Phi \in \mathbb{R}^{M \times N}$ with $M \ll N$ is the compressed measurement matrix, and estimates $\hat{x}$ from measurement $y \in \mathbb{R}^M$ by solving the ill-posed inverse problem.

In the optical encoder phase, typical CASSI utilizes binary random coded apertures with small correlation between wavelengths. Some approaches have been proposed to optimize the coded aperture from different perspectives [9], [10], [11]. Unfortunately, these methods have shown suboptimal performance due to their inability to align with the sparsity distribution under target content. It is a challenging task solely for reconstruction algorithm to recover the high-frequency information that has already been lost. As the visualization in Fig. 1 (a) (b), we demonstrate the above concerns with two simulation experiments. *Mask A* outperforms *Mask B* in perceiving the first scene, whereas

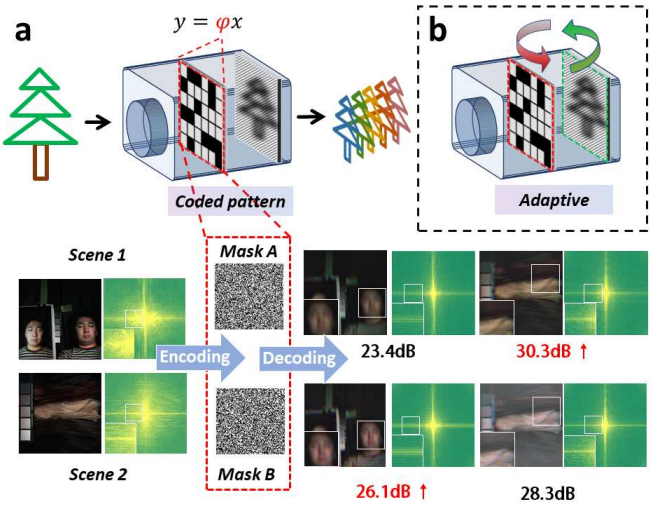


Fig. 1. Problem statement: (a) Spectral SCI system with fixed mask leads to unstable results: mask A and B represents two different random coded pattern, datasets are selected from CAVE [12] and decoded by TwIST algorithm [13], results demonstrate synthesized RGB images and FFT spectrum. (b) Self-adaptive SCI system is capable of adapting mask for specific scene.

the opposite result is found under the second scene. This discrepancy arises from different responses elicited by the fixed optical transfer function, which can be observed in the frequency spectrum where high-frequency information is lost to varying degrees, resulting in the distortion of spatial and spectral details.

The programmable digital micromirror device (DMD) or spatial light modulator (SLM) enables the imaging system to modify the encoded patterns during the acquisition phase. Several recent studies have demonstrated the superior performance of content-aware adaptive masks [14], [15], [16], [17], [18]. For example, Galvis *et al.* [15] used side information to adaptively sense high spatial frequency components

• *Zhan Shi, Hao Ye, Tao Lv, Yibo Wang, and Xun Cao are with the School of Electronic Science and Engineering, Nanjing University, Nanjing, 210023, China (email: {zhanshi, yehao, lvtao, ybwang}@smail.nju.edu.cn, caoxun@nju.edu.cn). Corresponding author: Xun Cao.*

for better reconstruction. Also, Saragadam *et al.* [18] introduced RGB-guided super-pixel segmentation for optimizing uniform non-overlapping sampling strategy proposed in [19]. However, these methods are often limited to hybrid systems, with an additional guide RGB camera to provide prior spatial information, which is inevitably restricted by calibration and systematic complexity. In the context of non-hybrid optical systems, a straightforward approach involves first reconstructing the acquired low-resolution signal and subsequently extracting physical prior information from it, but the reconstruction process is frequently a time-intensive endeavor that generates an approximation of the original signal. Inspired by the direct inference on compressed measurements [20], [21], the ultimate prior properties for adapting content-aware sampling strategy can be determined directly in the compressed domain (i.e. from the previous low-resolution shot image).

Besides, existing methods only manually adjust sampling strategy using physics-based heuristics, and then design corresponding reconstruction process. However, these isolated physical cues may not be decoded perfectly when solving the inverse problem with hard-interpretable but remarkable deep-learning models [22], [23], [24], [25]. Motivated by the recent deep-optics and task-specific camera exploration [3], [26], [27], adaptive optics optimized with the vision algorithm enabling fitting their intrinsic connection in a data-driven approach.

To this end, we propose a novel **C**ompact **S**elf-adaptive optical **C**oding framework for **S**pectral **C**ompressive **S**ensing, dubbed **3CS**, which adaptively optimizes coded patterns in the compressed domain without requiring an extra guide camera, which is suitable for single-path SCI systems (e.g. SD-CASSI [8] and DD-CASSI [7]). The "Adaptive Encoder + Digital Decoder" framework consists of an Initialized Module (IM), a lightweight Adaptive Encoder Predictor (AEP), and the E2E decoder. We define the distribution of the mask as an union set of a binary random distribution and an adaptive Content-aware Spatio-Spectral Variation (CSSV) distribution. The learnable parameters of the lightweight AEP are co-trained with the decoder network to extract high-frequency prior information from the t-1 compressed measurement, which helps acquire the high-frequency component CSSV distribution and optimize the mask at time t for better accuracy. In the digital decoder phase, the E2E decoder executes single-shot reconstruction. To be highlighted, our framework maintains snapshot capability in the sense that each final reconstruction is based on a single snapshot measurement, because fusing both previous and current measurements during reconstruction may leading drastic errors caused by motion artifacts. In this specific way, the adaptive CSSV distribution provides a gain for scene reconstruction and avoids steep errors caused by reduced inter-frame correlation when the scene undergoes significant motion.

In summary, we propose a novel self-adaptive coding framework to optimize HSI sensing and reconstruction, our main contributions are as follows:

1) This framework design is suitable for compact single-path SCI systems and various reconstruction models. The data-driven AEP and single-shot HSI decoders are optimized in an E2E manner, learning scene-adaptive parameters based on different imaging principles and decoding models.

2) AEP is capable of inferring the CSSV distribution of high-frequency priors, and updating the mask by uniting it with a random distribution. This setting exhibits high light efficiency and robustness to dynamic scenes with rapid motion.

3) Experiments are conducted on both static HSI datasets and dynamic hyperspectral videos. The results and model discussions validate the superiority of our approach for spectral video photography.

2 Related Works

2.1 Spectral SCI Camera Designs

Spectral SCI cameras typically achieve hyperspectral imaging through paradigm-specific optical modulation and by solving a sparsity-constrained inverse problem. Among them, coded aperture methods [8], [28], [29], which involve the utilization of coded apertures that selectively block incident light in the intermediate image plane and a dispersive optical element, then reconstruct 3D HSIs from 2D compressed measurements. Additionally, other system designs, such as diffraction-based [30], [31] and filter-array-based [32], [33] methods also employ similar paradigms of forward modulation and backward solving to construct specific forms of compressed measurements based on different principles.

Meanwhile, some hybrid camera designs [19], [34], [35], [36] enhance the performance of hyperspectral reconstruction by multiplexing different forms of data sampling through parallel optical paths. However, the temporal synchronization and spatial alignment between cameras create additional challenges for system calibration, while also placing higher demands on the alignment between optical elements.

Iterative algorithms [37], [38], [39], used in the inverse process of the system, which rely on time-consuming coding processes and hand-crafted prior knowledge, exhibit drawbacks in terms of performance and generality. To address these challenges, recent efforts have focused on using E2E deep-learning models [22], [23], [24], [25], [40]for HSI reconstruction. This approach has yielded impressive performance and demonstrated potential for optimizing complex restoration problems across a range of snapshot imaging systems.

For CASSI, the reconstruction performance is sensitive to both the coded aperture and the reconstruction algorithm utilized. Inspired by deep optical imaging [3], [26], [27], our framework unifies optical encoding and reconstruction decoding to pursue improved performance.

2.2 Mask Optimization Methods

The design of the mask plays a crucial role in determining the imaging sensing matrix, which directly impacts the reconstruction accuracy. Therefore, coded aperture optimization has gained increasing attention in recent years [9]. Several methods have been proposed to optimize the coded aperture based on the Restricted Isometry Property (RIP) in CS theory [10], [37], [41], [42]. Additionally, SS-CASSI [29] has been proposed, which utilizes color masks

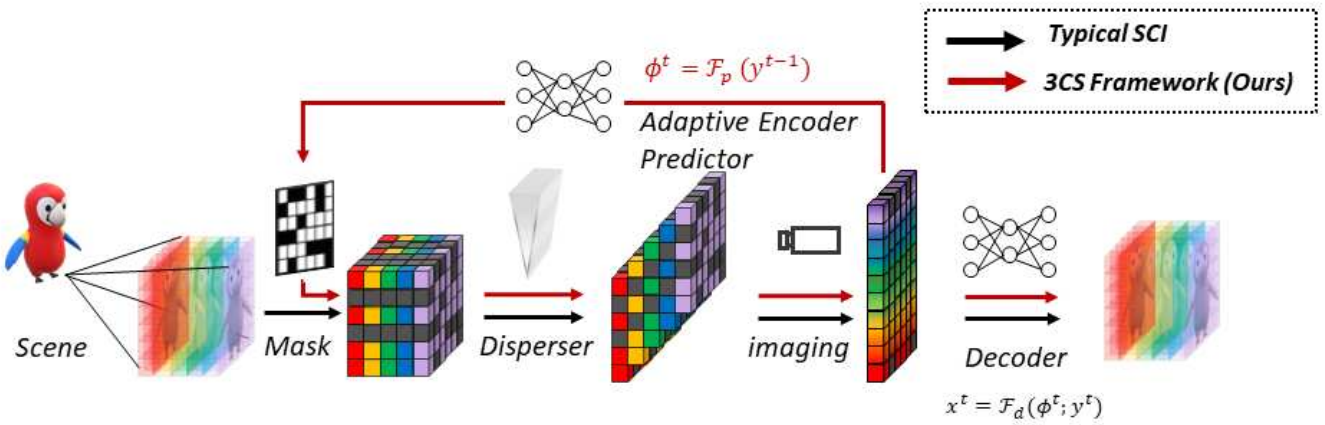


Fig. 2. Typical spectral SCI schematic (black arrows) and 3CS framework (red arrows)

for simultaneous spatio-spectral dual-domain modulation of HSI.

However, the constrained representation capacity of manually specified priors imposes limitations on their performance. Inspired by the principles of deep optics, some researchers have explored joint mask optimization and image reconstruction schemes, achieving improvements over isolated optimization. HyperReconNet [43] unified the coded aperture parameters and reconstruction model by embedding the entities of the repeated coded apertures as network weights on the basis of convolutional neural network. On the other hand, HerosNet [44] unfolded the optimization iterative process based on a deep unfolding network and jointly learned the binary optimal mask for recovering high-quality HSI.

2.3 Content-aware Adaptive Coding

Multiple recent research efforts [14], [15], [16], [17], [18] have provided evidence for the exceptional capabilities of adaptive masks that are tailored to the content of the imaging scene.

Ma *et al.* [14] used an RGB image to generate SLM sampling patterns in real-time, and then spread information from adjacent frames according to temporal correlation. However, correcting temporal errors in the presence of rapid motion is challenging. SASSI [18] utilized the same hardware configuration to perform scene-adaptive spatial sampling for the next frame. Its sampling strategy is guided by the super-pixel segmentation from the high-resolution RGB image. The mask settings for both of these works are based on the PMVIS paradigm [28], where the dispersion between sampling points is non-overlapping. Furthermore, other works also utilized information provided by RGB cameras from different perspectives for adaptive mask generation.

In contrast to these works, we will explore a more compact design by directly extracting prior information from low-resolution compressed images instead of using a hybird RGB system.

3 Methodology

3.1 Conventional SCI Model

Let $x(i,j,\lambda)$ denote the 3D spectral density of the target scene, where i,j index the spatial coordinates and λ indexes the spectral. Following the optical path in SD-CASSI, as shown in Fig 2, the encoded 2D measurement $y(i,j)$ compressed on the detector array can thus be formulated as

$$\begin{aligned} y(i,j) &= \int_{\Lambda} \Big(\iint_{V} \delta(u - i - \alpha(\lambda - \lambda_c))\delta(v - j) \cdot \\ &\qquad f(u,v,\lambda) c(u,v,\lambda)\, du\, dv \Big)\, d\lambda \\ &= \int_{\Lambda} x(i + \alpha(\lambda - \lambda_c), j, \lambda) c(i + \alpha(\lambda - \lambda_c), j) d\lambda \end{aligned} \tag{1}$$

where the incident light x is firstly modulated spatially by physical coded aperture (i.e., occlusion mask) $c(x,y)$. α is the coefficient determined by the degree of dispersion. After that, modulated spectral image is shifted by optical dispersion units (i.e., prism or gratings) along one spatial direction, where $\iint_V$ is the shifted spatial range and Λ is the spectral range. Finally, the imaging sensor captures 2D compressed measurement y that contains spectrally various coded mixed information.

We discretize the incident spectral as 3D HSI cube $X \in \mathbb{R}^{H \times W \times L}$, where H, W, and L indicate the HSI's height, width, and number of spectral channels, respectively. Then the discrete SD-CASSI forward model is

$$Y(m,n) = \sum_{k=1}^{L} T(X(:,:,k) \odot C(:,:,k)) \tag{2}$$

where $Y \in \mathbb{R}^{H \times (W + (L-1)*\Delta)}$ denotes discrete measurement and Δ denotes the shifting interval. $\odot$ is element-wise Hadamard product. $C \in \mathbb{R}^{H \times W \times L}$ denotes discrete 3D coded pattern. $T(\cdot)$ is the shifting operation similar to above continuous model. $\sum_{k=1}^{L}$ is summation operation represents imaging process that integrates the coded and shifted spectral data cube along the spectral dimension.

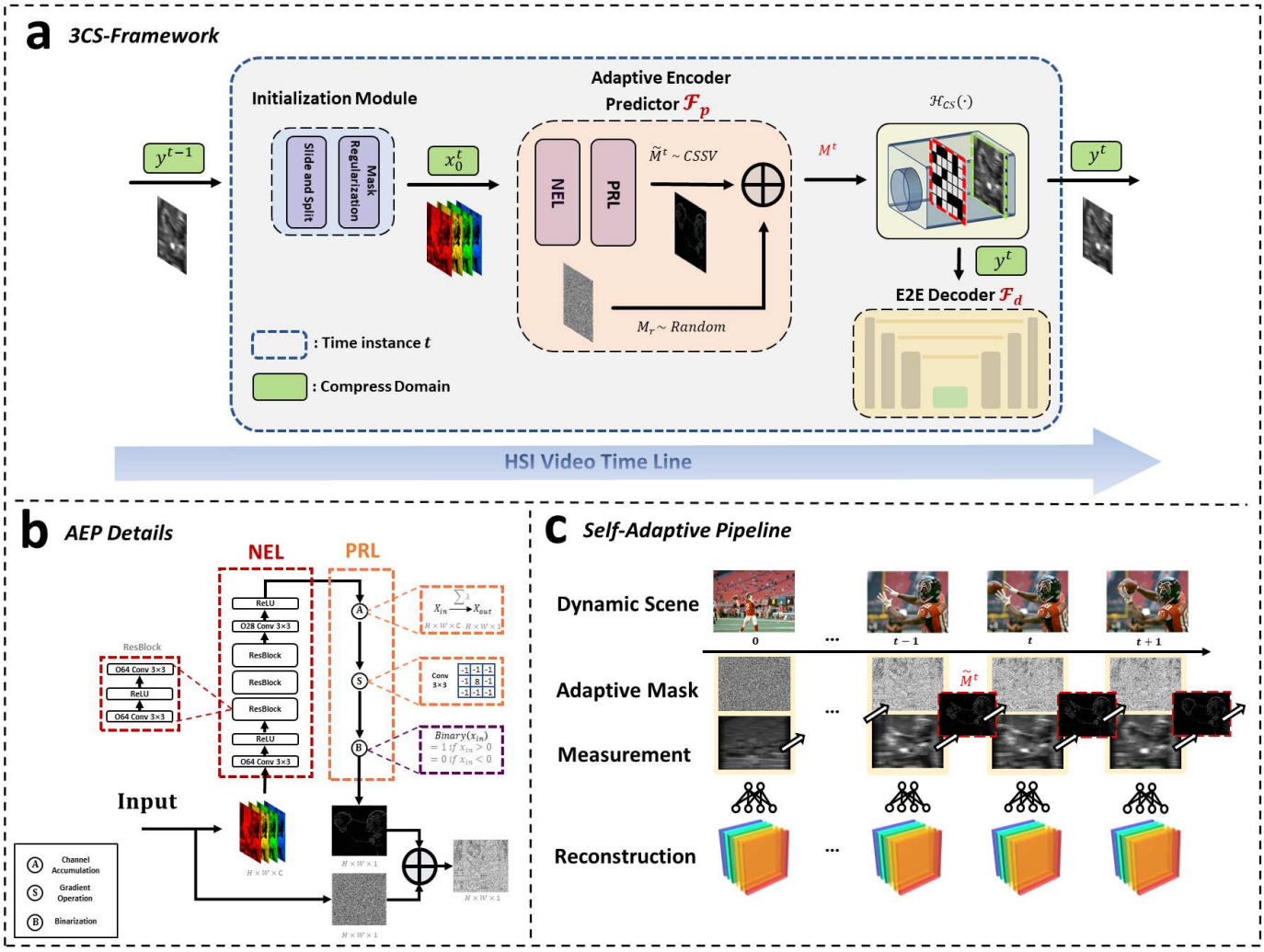


Fig. 3. Methodology: (a) details of 3CS framework. (b) details of AEP. (c) 3CS self-adaptive pipeline.

To illustrate the proposed model, we convert it to matrix form to obtain a subsampling structure consistent with the compressed sensing:

$$\hat{x} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\Phi\mathbf{x} - \mathbf{y}\|_2^2 + \lambda R(x) \quad s.t. \quad y = \Phi x + \epsilon \quad (3)$$

where $y \in \mathbb{R}^{H(W+(L-1)*\Delta)}$ and $x \in \mathbb{R}^{HWL}$ is vectorization terms of measurement Y and HSI cube X, $\Phi \in \mathbb{R}^{H(W+(L-1)*\Delta)\times HWL}$ denotes the sensing matrix of forward model, $\lambda R(x)$ is a prior regularization term.

Traditional E2E deep-learning methods embeds the regularization term $R(x)$ and noise ϵ into a nonlinear process, modeling that desired spectral signal x is the output of a decoder neural network, $\hat{x} = \mathbf{F}_d(y)$, d are the network parameters to be learned from the training data. Thus the optimization task suggest to solve:

$$\min_d \|x - \mathbf{F}_d(y)\|_2 \quad (4)$$

3.2 The Proposed 3CS Framework

The overall network architecture and internal module details of our proposed 3CS framework are shown in Fig. 3. In this subsection, we elaborate on the system adaptive pipeline, the details of the Initialization Module (IM), the Adaptive Encoder Predictor (AEP), and our joint optimization strategy together with E2E decoders in light of previous descriptions.

Self-Adaptive Pipeline. Starting from equation(3), to address the content-aware self-adaptive coding problem, we propose a new timing model for the spectral SCI schematic:

$$y^t = \Phi(M^t)x^t + \epsilon \quad (5)$$

Our model considers dynamic sampling strategy by updating the sensing matrix Φ via adaptive mask M^t along frames.

For time t, the prior information of the scene is provided by the previous shot image or reconstruction results. In order to ensure temporal correspondence between prior and current target scene, we only utilize the last frame, which can mitigate disalignment errors caused by rapid motion in the scene.

Furthermore, inspired by compressed learning, whole-process reconstruction of HSI is not a necessary process for optimal mask inference. Instead, we are only interested in certain properties and can effectively infer them directly from compressed domains.

Fig. 3 (c) illustrates the pipeline of our strategy when processing HSI video sequences. For time t, we determine the optimized mask from unreconstructed image acquired at time $t-1$ by AEP $\mathbf{F}_p$:

$$M^t = \mathbf{F}_p(y^{t-1}) \tag{6}$$

Hardware device such as DMD or SLM is first altered to adapt masks as M^t, then the camera acquire modified compressed HSI y^{t-1} within exposure time:

$$y^t = \Phi(\mathbf{F}_p(y^{t-1}))x^t + \epsilon \tag{7}$$

Finally, the HSI is reconstructed by a decoding neural network from current shot image to avoid erroneous, instead of fusing multiple frames:

$$\hat{x}^t = \mathbf{F}_d(y^t) \tag{8}$$

We note here that the process $M^{t+1} = \mathbf{F}_p(y^t)$ and $\hat{x}^t = \mathbf{F}_d(y^t)$ go simultaneously and independently, avoid time-consuming interaction for waiting.

Initialization Module (IM). Acquired measurement $Y \in \mathbb{R}^{H\times\ (W+(L-1)*\Delta)}$, the objective of the initialization module is to initial the 2D measurement into a 3D HSI $X_0 \in \mathbb{R}^{H\times\ W\times L}$, wherein W and H represent the spatial dimensions of the frames, and L denotes the number of spectral channels. To achieve this, a sliding window is utilized to crop the measurement Y subsequently in increments of Δ to generate L HSI frames. The resulting L frames are ultimately concatenated along the channel dimension, culminating in the formation of a 3D HSI. To address the upsampling errors, each channel is multiplied with the coded pattern to constrain them, as these points have been completely obstructed during the encoding process:

$$X_0(m,n,k) = Y(m+k*\Delta,:) \odot C(:,:,k) \tag{9}$$

Adaptive Encoder Predictor (AEP). Different from SASSI [18] that utilized sparse sampling pattern, we follow the dense mask format of CASSI. Rather than modeling the entire mask distribution, we explicitly decompose the adaptive mask as a known constant binary random mask M_r, which obeys the binary random distribution, merges an Content-aware Spatio-Spectral Variation (CSSV) distribution mask $\tilde{M^t}$:

$$M^t = M_r \oplus \tilde{M^t} \tag{10}$$

Here $\oplus$ is element-wise union operation, $\tilde{M^t}$ is the adaptive component target high-frequency priors based on the scene, which will be inferred by AEP based on the compressed domain information of the last frame.

Thus, for the optimized compressed sensing image, it consists not only of fundamental randomly encoded sampling components, but also of continually-updated high-frequency context uncertainty part from last frames. Note that the process is not linear because there are many sampling points that exist simultaneously in both parts of the mask.

This increment can provide a gain for the reconstruction of high-frequency information. Eq. 10 implicitly embeds the CSSV distribution into the random mask. Even when sampling locations in CSSV are off from the desired pattern, the mask still maintains its randomness. By defining in this twofold way, we avoid steep errors due to mistake sampling caused by rapid motion, in that the compressed image still maintains a large degree of dense random encoding components.

For AEP, the key task is to learn a good representation of the high-frequency details. As shown in Fig. 3(b), AEP consists of Neural Embedding Layer (NEL) and Physical Regulation Layer (PRL). In order to extract prior features A from the initialized compressed domain HSI X_0, we first adopt NEL as the feature extractor to evaluate the desired components maps:

$$A = \mathbf{F}_{NEL}^{L->L}(X_0) \tag{11}$$

The initialized cube X_0 will be processed by a convolution layer, three residual blocks and another convolution layer without channel down-sampling. Then the feature maps A will be sent to the PRL to estimate the parameters of binary CSSV mask.

PRL mainly includes three steps: First, we perform channel accumulation operation to fusion the high-dimensional features extracted by NEL along λ dimension into a 2D spatio-spectral representation to align with the mask sampling position; then, we convolve the low-dimensional feature with gradient operators to extract the high-frequency information of the spatio-spectral variation features; finally, we design an element-wise binarization function to transforms the continuous feature matrix into the mask form through binarization operators $Binary(\cdot)$:

$$\tilde{M}_t = \mathbf{F}_{PRL}^{L->1}(A) = Binary(\sum_{\lambda} A * \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}) \tag{12}$$

E2E Decoder and Joint Optimization. In the reconstruction stage, we utilize an E2E deep network for decoding. As the parameters of AEP are trained jointly with the reconstruction network, the way it infers the distribution of high-frequency information is influenced by the reconstruction network. To analyze this effect, we employ several mainstream models with different structures, which will be detailed in the experimental section.

The adaptive encoder predictor network $\mathbf{F}_p$ and HSI decoder network $\mathbf{F}_d$ are optimized jointly for the same attempt for higher fidelity reconstruction. All the parameters are optimized by minimizing the Root Mean Square Error (RMSE) loss function through E2E training in Eq. 13:

$$\begin{aligned} p,d &= \min_{p,d} \left\| x^t - \mathbf{F}_d(\Phi(\mathbf{F}_p(y^{t-1}))x^t) \right\|_2 \\ &= \min_{p,d} \left\| x^t - \mathbf{F}_d(\Phi(\mathbf{F}_p(\Phi^{t-1}x^{t-1}))x^t) \right\|_2 \end{aligned} \tag{13}$$

Due to the difficulty in predicting the future distribution of videos and the unavailability of suitable hyperspectral video training data, we assume in the formulation that x^{t-1} and x^t are the same during the training process. This allows the network to learn and fit better sampling patterns for known scenes using HSI datasets. Meanwhile, in order to reduce the complexity of the problem, we only optimize the process between the first frame and the second frame (i.e. from random mask to optimized mask), this is because

the binarization operation would effectively suppresses the artifacts introduced to subsequent frames.

$$p, d = \min_{p,d} \frac{1}{D} \sum_{n=1}^{D} \|x_n - \mathbf{F}_d(\Phi(\mathbf{F}_p(\Phi(M^r))x_n))x_n)\|_2 \quad (14)$$

where $\mathcal{D}^t$ denotes the total number of the training HSIs, $\Phi(M^r)x_n$ represents the measurement of scene x_n captured by random mask.

The deep reconstruction network is known to be highly sensitive to mask variations, which introduces challenges during the training process. As the parameters of the AEP network change, the predicted mask also undergoes changes. Consequently, for each iteration, the input to the decoder network for the same scene differs, leading to difficulties in achieving convergence during co-training.

To facilitate the convergence of the model during training, we designed a special two-stage optimization (illustrated in Algorithm 1. First, the scene is directly passed through the PRL layer to obtain the CSSV distribution under the condition of known ground truth, and the optimized mask is obtained using Eq. 10 (here NEL does not perform inference, and the mask is fixed). We pre-trained the decoding network on the optimized mask and the training set, so that the decoding network will first learn a local optimal solution for the CSSV distribution. Finally, we optimize the joint model together using Eq. 14 to obtain the final convergence result.

Algorithm 1: *3CS* Joint Optimization

Input: $\mathcal{D}^t$, $\mathcal{D}^v$, M_r, initialized p, d
Output: $\hat{p}, \hat{d}$
Generating $M^{pre} = M_r \oplus F_{PRL}(x_d), x_d \in \mathcal{D}^t$
Pre-train $F_{d'}(\cdot)$ on $(M^{pre}; \mathcal{D}^t)$;
Initialize joint model with p, d';
while *not converge* **do**
  for $i = 1, ..., t$ **do**
    train $F_{\hat{p}}(\cdot), F_{\hat{d}}(\cdot)$ on $(M_r; \mathcal{D}^t)$;
  valid $F_{\hat{p}}(\cdot), F_{\hat{d}}(\cdot)$ on $(M_r; \mathcal{D}^v)$;
Return result;

4 Experiment

To verify the effectiveness of the proposed method, we conduct experiments on the commonly available synthetic datasets. We demonstrate performances improvement between popular E2E baseline methods and 3CS adaptive setup. Finally we will analyze advancedness with the light-efficiency and dynamic motion robustness of our method. Both quantitative and qualitative perspectives are employed to investigate above concern.

4.1 Datasets

CAVE: The CAVE [12] database is comprised of 32 HSIs including a diverse range of real-world materials and objects. These images were captured under standard D65 illumination, using a tunable filter and a cooled CCD camera, resulting in a spatial resolution of 512 x 512 pixels and 31 spectral bands spanning the wavelength range of 400-700nm, with a 10nm interval.

KAIST: The KAIST [45] HSIs dataset owns a high spatial resolution of 3376x2704, which was captured by employing tunable filters and the Pointgrey Monochromatic camera GS3-U3-91S6M-C. Similar to the CAVE dataset, the KAIST dataset also encompasses 31 bands, which covers the spectral range between 420 and 720 nm with a step size of 10 nm.

Dataset Details: We follow the dataset settings provided in [23], which randomly select $256 \times 256 \times 28$ HSIs from CAVE as training sample and set 10 scenes from KAIST for model testing. The 28 spectral bands ranging from 450 to 650nm were obtained through interpolation. Measurements are computed by simulating the compressing process Φ of SD-CASSI optical system [8] with 256×256 spatial mask modulation and then dispersed Δ step is set to two-pixel.

4.2 Evaluation Metrics

We adopt metrics of Peak Signal to Noise Ratio (PSNR), Mean Square Errors (MSE) and Structured Similarity Index Metrics (SSIM) [46] for evaluating quantitative reconstruction quality. Note that unlike PSNR and SSIM, MSE is inversely proportional to the reconstruction quality. Typically, PSNR (dB), SSIM $\in [0, 1]$, and MSE $\in [0, +\infty)$ are calculated per-band and averaged overall bands.

For qualitative analysis, we will demonstrate the pseudo-color image of spectral bands, RGB color images synthesized from HSIs, and their locally magnified images.

4.3 Implementation Details

Comparison Methods. In this study, our experiments primarily validate the effectiveness of the proposed 3CS adaptive framework in terms of single-shot decoder accuracy and light efficiency compared to E2E methods alone without adaptive mask, (i.e. with a fixed real mask). Among the representative E2E reconstruction algorithms, we selected λ-Net [22], TSA-Net [23], DGSMP [24], and HD-Net [25], which have different design emphases. We will demonstrate the general applicability of the proposed method on these reconstruction models. Furthermore, we select HD-Net as decoder example to process two hyperspectral video reconstruction, demonstrating the robustness of our method against temporal correlation reduction when processing dynamic spectral videos.

Parameter Settings. We jointly training AEP $\mathbf{F}_p$ and E2E decoder $\mathbf{F}_d$ by minimizing RMSE Loss in Eq. 14 by optimizing Algorithm 1. The joint model is optimized by ADAM optimizer with setting betas=(0.9, 0.999) and learning rate of 4×10^{-4} with periodic decline. We train the network using Pytorch and two NVIDIA GeForce RTX 3090 GPUs.

4.4 Performance Comparision

Table 1 evaluates and compares the reconstruction results on the 10 test scenes in KAIST. We select four representative algorithms as decoder $\mathbf{F}_d$. For every decoder model, we compare PSNR and SSIM between E2E baseline framework and our proposed 3CS adaptive framework.

TABLE 1
Reconstruction results (PSNR (dB) / SSIM) for single-shot reconstruction from E2E and 3CS framework.

Decoder Model	λ-Net		TSA-Net		DGSMP		HD-Net	
Framework	E2E Baseline	3CS Adaptive	E2E Baseline	3CS Adaptive	E2E Baseline	3CS Adaptive	E2E Baseline	3CS Adaptive
Scene01	32.19 / 0.8963	**33.69 / 0.9145**	33.10 / 0.9049	**33.37 / 0.9145**	33.44 / 0.9241	**34.24 / 0.9390**	33.27 / 0.9153	**34.47 / 0.9335**
Scene02	30.28 / 0.8458	**32.82 / 0.8923**	31.95 / 0.8898	**32.91 / 0.9065**	33.12 / 0.9220	**34.42 / 0.9397**	32.86 / 0.9003	**34.78 / 0.9309**
Scene03	32.08 / 0.9265	**34.27 / 0.9411**	32.59 / 0.9255	**32.98 / 0.9301**	32.65 / 0.9328	**33.27 / 0.942**	33.83 / 0.9275	**35.19 / 0.9469**
Scene04	39.19 / 0.9658	**40.83 / 0.9740**	38.57 / 0.9671	**39.82 / 0.9727**	36.45 / 0.9680	**37.87 / 0.9692**	38.24 / 0.9656	**39.39 / 0.9738**
Scene05	29.34 / 0.8898	**30.99 / 0.9191**	30.29 / 0.9092	**30.56 / 0.9195**	29.86 / 0.9230	**30.78 / 0.9380**	31.21 / 0.9216	**32.25 / 0.9412**
Scene06	29.83 / 0.8970	**31.98 / 0.9218**	31.73 / 0.9230	**32.99 / 0.9382**	32.83 / 0.9412	**34.58 / 0.9563**	32.26 / 0.9282	**34.60 / 0.9563**
Scene07	29.01 / 0.8645	**31.62 / 0.9023**	30.65 / 0.8956	**31.27 / 0.9088**	30.94 / 0.8996	**31.63 / 0.9179**	31.41 / 0.8977	**33.34 / 0.9280**
Scene08	28.21 / 0.8817	**30.70 / 0.9097**	30.01 / 0.9144	**31.38 / 0.9344**	30.82 / 0.9310	**32.57 / 0.9485**	30.18 / 0.9132	**32.51 / 0.9457**
Scene09	29.55 / 0.8881	**33.30 / 0.9166**	31.24 / 0.9178	**31.93 / 0.9282**	30.32 / 0.9174	**31.37 / 0.9399**	32.33 / 0.9192	**34.27 / 0.9409**
Scene10	28.09 / 0.8392	**29.49 / 0.8771**	29.55 / 0.8899	**30.83 / 0.9120**	31.04 / 0.9335	**32.56 / 0.9518**	30.25 / 0.9011	**31.72 / 0.9374**
Average	30.77 / 0.8895	**32.97 / 0.9168**	31.97 / 0.9137	**32.80 / 0.9265**	32.15 / 0.9293	**33.33 / 0.9442**	32.58 / 0.9190	**34.25 / 0.9435**

As highlighted in red-bold, results depict that our 3CS attains consistent and significant improvement as compared to traditional E2E methods on all scenes. Taking HD-Net decoder as an example, our method achieves an average of 1.67dB increase in PSNR and an average of 0.0245 increase in SSIM, reaching 34.25dB and 0.9435 respectively. For λ-Net, 3CS adaptive framework achieves an average of 2.2dB and 0.0273 improvement. The proposed method also achieves 1.18dB/0.0149 gains on DGSMP. But for TSA-Net, we achieve little gain of 0.83dB and 0.0128. It demonstrates that the proposed 3CS can better perceive the spatio-spectral infomation from adaptive encoder and achieve more fidelity reconstruction by E2E decoder.

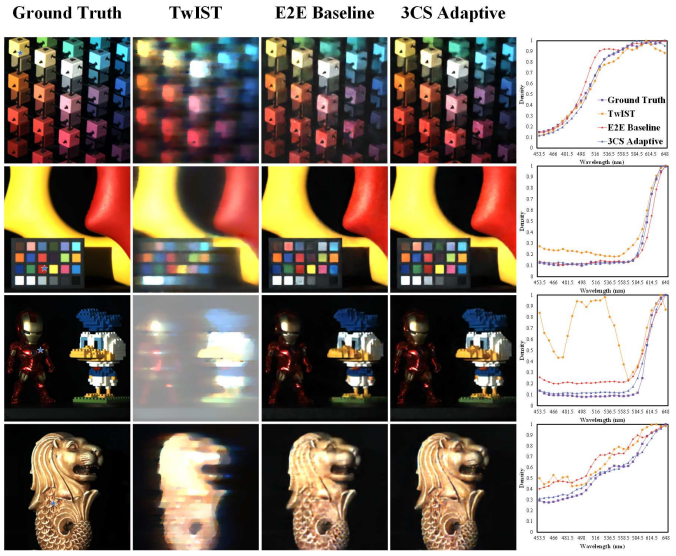


Fig. 4. Synthesized RGB images and selected spectral curves for the ground truth and decoding results from the TwIST iteration, E2E methods and our 3CS framework. Scenes are selected from the test set.

Fig. 4 shows a visual example of RGB images synthesized from HSIs. The spatial details and color rendering can to some extent reflect the accuracy of spatial and spectral reconstruction. From the figure, it is evident that the E2E network is inferior to the TwIST iterative algorithm. Note

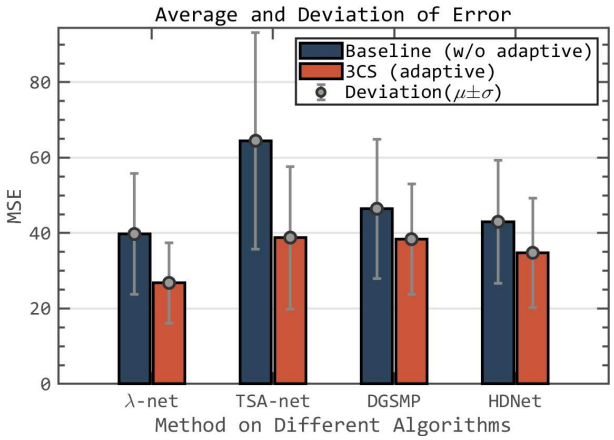


Fig. 5. Average MSE and standard deviation on 10 test scenes

that in the figure, the E2E model uses HD-Net, DGSMP, TSA-Net, and λ-Net from top to bottom, respectively. Moreover, by incorporating our adaptive framework, the model achieves more visually pleasing results in terms of both color and spatial detail fidelity. We believe this is due to the additional perception of high-frequency contextual information provided by the adaptive mask. Besides, it can be seen that the spectral curves generated by our method are more consistent with the reference.

As illustrated in Fig. 5, we also present the MSE and standard deviation among the ten testing results, which indicates that our method not only reduces the average error, but also narrows the deviation among the results. This finding partly validates our initial intention of enhancing the robustness of encoding for different scenes.

In Fig. 6, we also plot 5 reconstructed spectral channel images (out of 28) and their locally magnified details. we compared two frameworks using above mentioned 4 reconstruction algorithms with different scenes. This does not affect fairness since there was no direct comparison between different reconstruction algorithms. As for the different spectral channels, the 3CS exhibits more spatial details and clearer texture compared to the E2E.

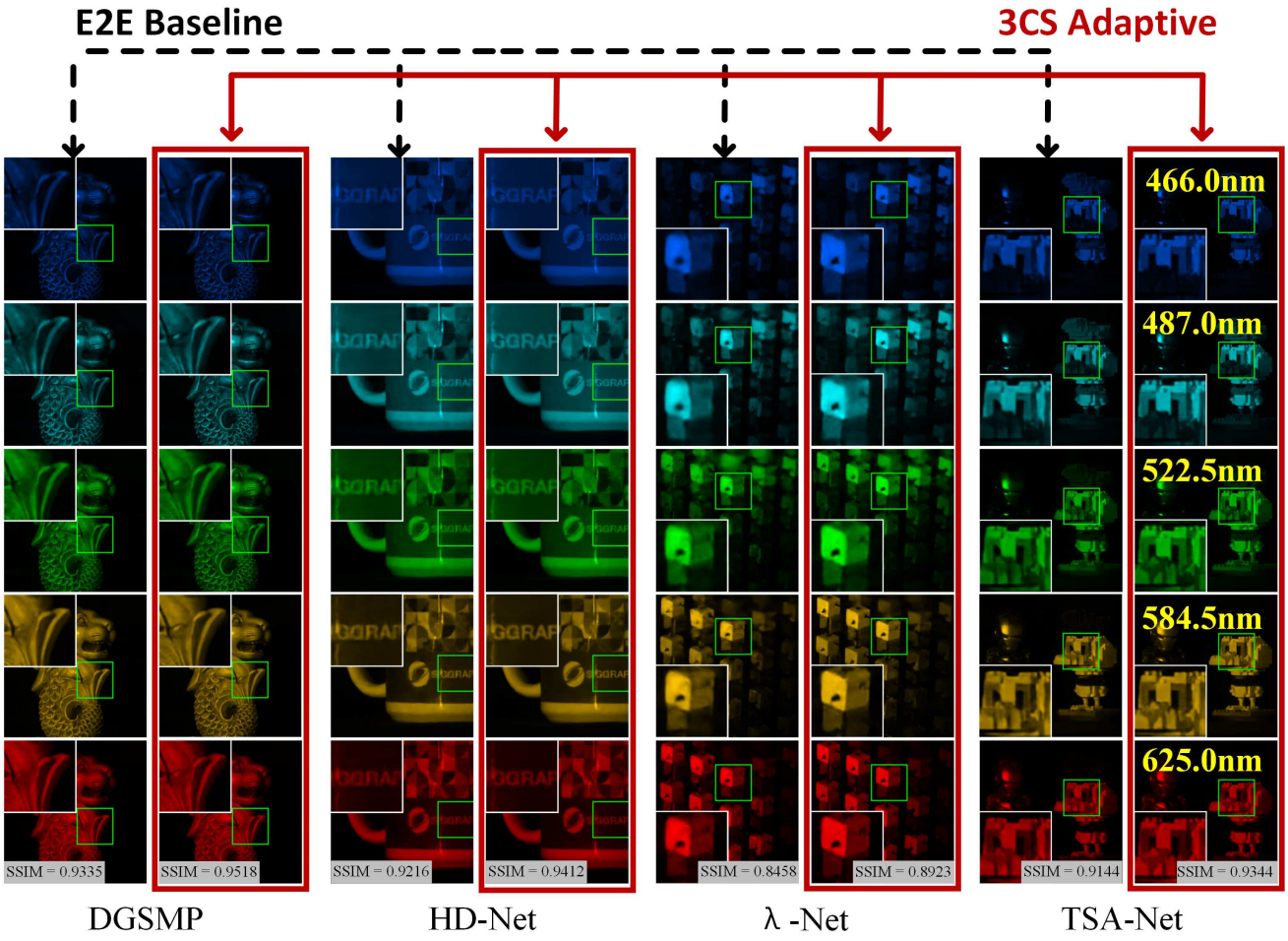


Fig. 6. Reconstructed images with 5 (out of 28) spectral channels using E2E and 3CS framework, each image is grayscale with pseudo-color. SSIM is averaged over all bands.

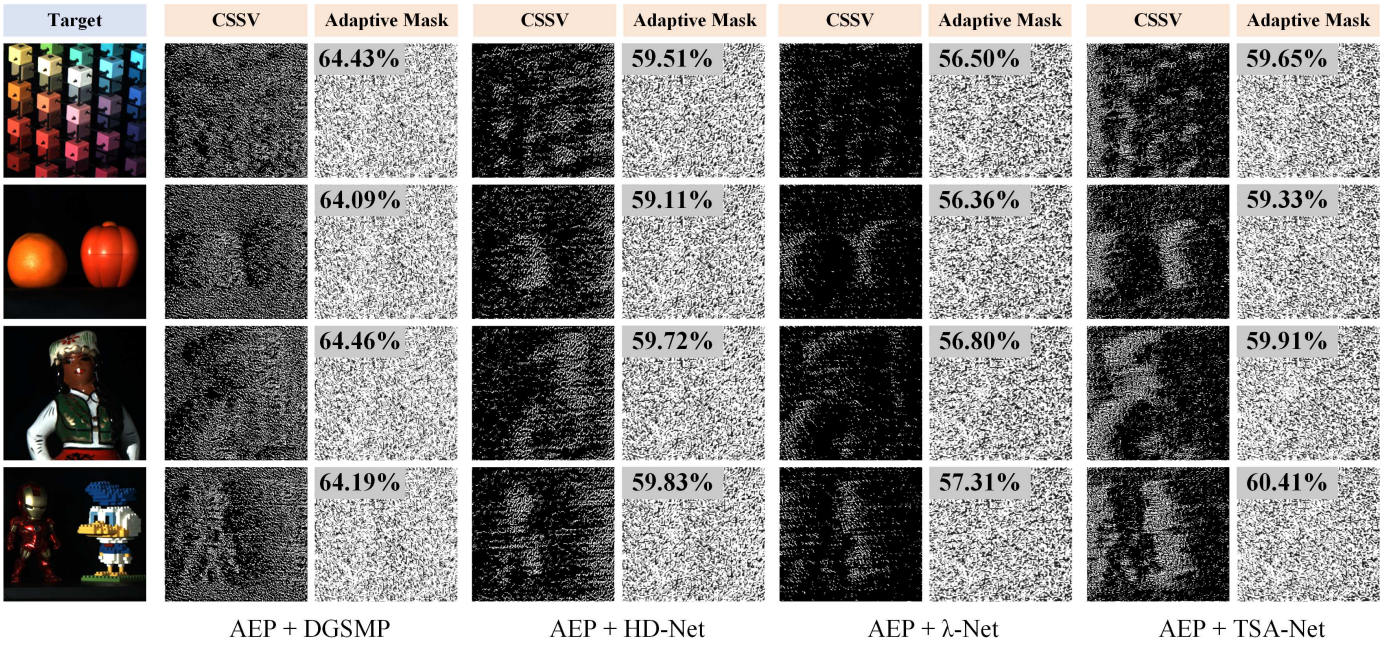


Fig. 7. CSSV distribution and adaptive mask predicted by AEP together with different decoders. The left panel of each pair shows the CSSV distribution, and the right panel shows the final generated mask. Additionally, the numerical values in the top left corner denote the light-throughput.

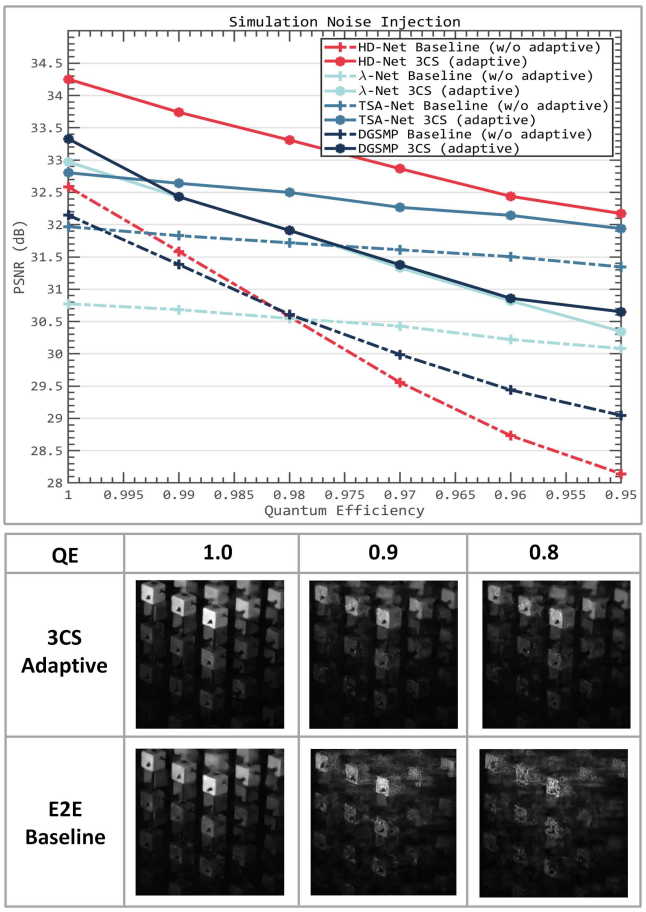


Fig. 8. Noise Analysis : (top) Average PSNR decline on 10 test scenes, (bottom) visual example on 522.5nm (decoded by HD-Net). All experiments are conducted assuming bit depth is 2048.

Fig. 7 presents the adaptive masks generated by AEP for four different scenes jointly optimized by various decoders, and the intermediate CSSV. Although the light-throughput values may vary, they generally improve much compared to 50% that of traditional random mask. It can be observed that the CSSV distribution is correlated with the semantic edges of the scene, thereby guiding the mask to preserve more high-frequency information. Further more, the AEP extracts different CSSV saliencies when joined with various decoders, which further validates our method's capability to enable the AEP to learn the suitable sampling patterns for the selected decoder.

4.5 Noise Robustness

The mask adapting method of the 3CS framework not only increases the high-frequency sampling components but also enhances the mask's light throughput. This setting enables the system to have more robustness to noise even when the sensor Quantum Efficiency (QE) is limited. However, noise also affects the mask prediction process. In this subsection, we conducted noise injection experiments to analyze the framework's effect on light-efficiency. Following [23], we model the shot noise considering QE and dynamic range as binomial distribution function:

$$y_{noise} = \mathcal{B}(y * d/QE, QE)/d \tag{15}$$

Here y is normalized noise-free measurement and y_{noise} is noise-injected measurement, d is bit depth decided by camera dynamic range, QE is from 0 to 1.

We both inject noise in E2E method and 3CS framework, which are both trained on clean data. Note that the 3CS is injected twice (i.e. t-1 measurent for mask prediction and t optimized measurement for reconstruction). Fig. 8 shows metric and visual results under different QE proving that our method maintains superior performance even under the influence of noise, while the E2E method is significantly affected. Obviously, without 3CS framework, the values of PSNR reconstructed by HD-Net drop substantially by 4.44dB. In contrast, with our adaptive strategy, it only decreases by 2.08dB.

4.6 Hyperspectral Video Evaluation

It stands to reason that the performance will degrade due to temporal artifacts between frames, even leading to a significant drop when processing scenes with rapid motion, because the correlation of the temporal signal may decrease significantly. We address this issue through the specific mask distribution: Eq. 10 implicitly embeds the CSSV distribution into the random mask. Even when the predicted sampling locations in CSSV are off from the desired pattern, the mask still maintains certain randomness. Besides, for every frame, 3CS execute single-shot reconstruction instead of fusing previous measurements, because fusing both measurement t-1 and t during reconstruction may cause drastic errors caused by motion artifacts. In this section, we conduct a series of experiments on simulated hyperspectral videos to validate the proposed approach and demonstrate its performance.

Due to the lack of available real hyperspectral videos, all the data in this section are of following settings: the motion of the scene is simulated by cropping a rolling window from the HSIs in the KAIST dataset with preset step value. As shown in the Fig. 9, we simulate a sequence of a bus in motion. At time $t = 0$, the mask is initialized with random values due to the lack of reference information. Since the mask is not adaptive, so the reconstructed results deviate only slightly from the baseline. From time $t \geq 1$ onwards, the mask starts to adaptively optimize based on the scene distribution of the previous frame. It can be observed that the CSSV distribution is correlated with the semantic edges of the scene, thereby guiding the mask to preserve more content-aware information. Moreover, even when the bus is in motion, the extracted CSSV distributions from consecutive frames still exhibit some level of correlation. The reconstruction quality of our method is improved and is comparable with ideal. Besieds, we apply the proposed method for reconstructing another dynamic spectral video in Fig. 10.

The above video experiments demonstrate typical dynamic scenes. To showcase the robustness against situation with unpredictable and dramatic motion, we conducted experiments on the 10 test scenes mentioned in Section 4.1 by rolling 100 pixels along the spatial dimension. Table 2 lists the performance of the model, the PSNR and SSIM demonstrate a minimal and reasonable decline of 1dB and 0.01, respectively, outperforming the baseline method still.

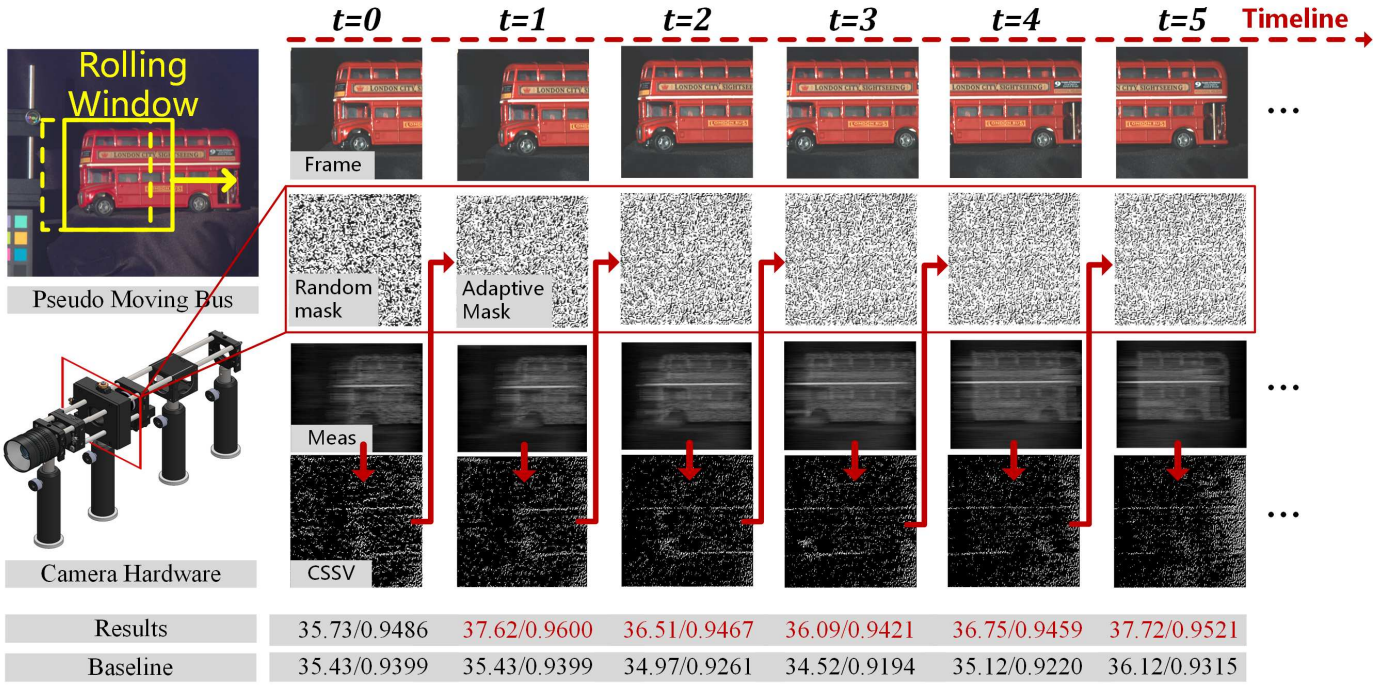


Fig. 9. Hyperspectral video reconstruction of the simulated moving bus (the origin HSI is resized to 622×777 and the moving step is set to 50 pixels)

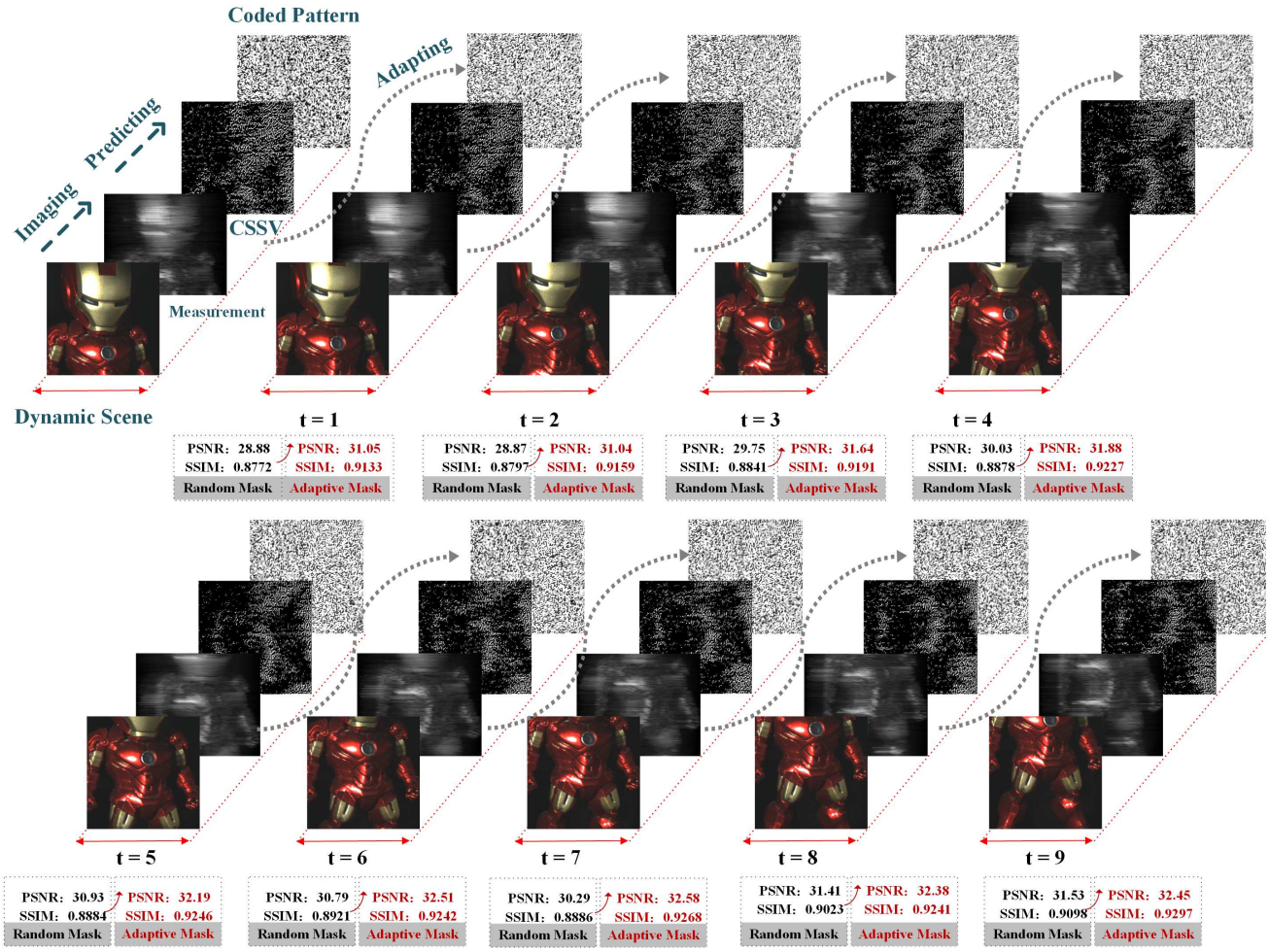


Fig. 10. Hyperspectral video of the simulated flying ironman (the origin HSI is resized to 1163×1452 and the moving step is set to 22 pixels)

TABLE 2
Decline in accuracy due to motion

Framework	E2E w/o Adaptive	3CS Adaptive (Move 0 Pixels)	3CS Adaptive (Move 100 Pixels)
λ-Net	30.77 / 0.8895	32.97 / 0.9168	32.81 / 0.9159
TSA-Net	31.97 / 0.9137	32.80 / 0.9265	32.65 / 0.9250
DGSMP	32.15 / 0.9293	33.33 / 0.9442	33.06 / 0.9427
HD-Net	32.58 / 0.9190	34.25 / 0.9435	34.14 / 0.9424

To provide evidence to show that the mask can be calculated within the internal of two frames, the inference time and the parameters of the AEP and Decoder are evaluated in the Table 3. The results demonstrate that the parameter count of the AEP is significantly smaller than that of the decoder (taking HD-Net as an example), with a difference of an order of magnitude. Moreover, it takes less than 8ms to calculate the mask between two frames, the speed can be improved with C++ implementation, making it an appealing approach for real-time video photography.

TABLE 3
Parameters and inference time of the AEP and decoder, based on the Intel i7-6700K CPU and NVIDIA GeForce RTX 1080 GPU

	Params (million)	Inference Time (s)
AEP	0.2379	0.00797
Decoder	2.3674	0.01037

5 Conclusion

In this paper, we propose and validate an innovative compact and adaptive coding framework for spectral SCI. The framework effectively extracts high-frequency information in the compressed domain by leveraging contextual cues and updates the physical mask to enhance perception and reconstruction capabilities for hyperspectral video photography. However, certain drawbacks, such as limited real-world experiments, still remain. In our future work, we will focus on improving the accuracy and interpretability of the model. Additionally, the framework will be implemented on a hardware prototype system for sufficient real-world validation.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62025108), the Leading Technology of Jiangsu Basic Research Plan (No. BK20192003), and the Key R&D Plan of Jiangsu Province (No. BE2022155).

References

[1] N. Gat, S. Subramanian, J. Barhen, and N. Toomarian, "Spectral imaging applications: remote sensing, environmental monitoring, medicine, military operations, factory automation, and manufacturing," in *25th AIPR Workshop: Emerging Applications of Computer Vision*, vol. 2962. SPIE, 1997, pp. 63–77.

[2] N. T. Clancy, G. Jones, L. Maier-Hein, D. S. Elson, and D. Stoyanov, "Surgical spectral imaging," *Medical image analysis*, vol. 63, p. 101699, 2020.

[3] J. Suo, W. Zhang, J. Gong, X. Yuan, D. J. Brady, and Q. Dai, "Computational imaging and artificial intelligence: The next revolution of mobile vision," *arXiv preprint arXiv:2109.08880*, 2021.

[4] N. Hagen and M. W. Kudenov, "Review of snapshot spectral imaging technologies," *Optical Engineering*, vol. 52, no. 9, pp. 090 901–090 901, 2013.

[5] X. Cao, T. Yue, X. Lin, S. Lin, X. Yuan, Q. Dai, L. Carin, and D. J. Brady, "Computational snapshot multispectral cameras: Toward dynamic capture of the spectral world," *IEEE Signal Processing Magazine*, vol. 33, no. 5, pp. 95–108, 2016.

[6] D. L. Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[7] M. E. Gehm, R. John, D. J. Brady, R. M. Willett, and T. J. Schulz, "Single-shot compressive spectral imaging with a dual-disperser architecture," *Optics express*, vol. 15, no. 21, pp. 14 013–14 027, 2007.

[8] A. Wagadarikar, R. John, R. Willett, and D. Brady, "Single disperser design for coded aperture snapshot spectral imaging," *Applied optics*, vol. 47, no. 10, pp. B44–B51, 2008.

[9] G. R. Arce, D. J. Brady, L. Carin, H. Arguello, and D. S. Kittle, "Compressive coded aperture spectral imaging: An introduction," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 105–115, 2013.

[10] A. Parada-Mayorga and G. R. Arce, "Colored coded aperture design in compressive spectral imaging via minimum coherence," *IEEE Transactions on Computational Imaging*, vol. 3, no. 2, pp. 202–216, 2017.

[11] N. Diaz, C. Hinojosa, and H. Arguello, "Adaptive grayscale compressive spectral imaging using optimal blue noise coding patterns," *Optics & Laser Technology*, vol. 117, pp. 147–157, 2019.

[12] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, "Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum," *IEEE transactions on image processing*, vol. 19, no. 9, pp. 2241–2253, 2010.

[13] J. M. Bioucas-Dias and M. A. Figueiredo, "A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Transactions on Image processing*, vol. 16, no. 12, pp. 2992–3004, 2007.

[14] C. Ma, X. Cao, R. Wu, and Q. Dai, "Content-adaptive high-resolution hyperspectral video acquisition with a hybrid camera system," *Optics letters*, vol. 39, no. 4, pp. 937–940, 2014.

[15] L. Galvis, D. Lau, X. Ma, H. Arguello, and G. R. Arce, "Coded aperture design in compressive spectral imaging based on side information," *Applied optics*, vol. 56, no. 22, pp. 6332–6340, 2017.

[16] X. Ma, H. Zhang, X. Ma, G. R. Arce, T. Xu, and T. Mao, "Snapshot compressive spectral imaging based on adaptive coded apertures," in *Compressive Sensing VII: From Diverse Modalities to Big Data Analytics*, vol. 10658. SPIE, 2018, pp. 11–19.

[17] H. Zhang, X. Ma, and G. R. Arce, "Compressive spectral imaging approach using adaptive coded apertures," *Applied Optics*, vol. 59, no. 7, pp. 1924–1938, 2020.

[18] V. Saragadam, M. DeZeeuw, R. G. Baraniuk, A. Veeraraghavan, and A. C. Sankaranarayanan, "Sassi—super-pixelated adaptive spatio-spectral imaging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 7, pp. 2233–2244, 2021.

[19] X. Cao, X. Tong, Q. Dai, and S. Lin, "High resolution multispectral video capture with a hybrid camera system," in *CVPR 2011*. IEEE, 2011, pp. 297–304.

[20] S. Lohit, K. Kulkarni, and P. Turaga, "Direct inference on compressive measurements using convolutional neural networks," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 1913–1917.

[21] A. Adler, M. Elad, and M. Zibulevsky, "Compressed learning: A deep neural network approach," *arXiv preprint arXiv:1610.09615*, 2016.

[22] X. Miao, X. Yuan, Y. Pu, and V. Athitsos, "l-net: Reconstruct hyperspectral images from a snapshot measurement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4059–4069.

[23] Z. Meng, J. Ma, and X. Yuan, "End-to-end low cost compressive spectral imaging with spatial-spectral self-attention," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*. Springer, 2020, pp. 187–204.

[24] T. Huang, W. Dong, X. Yuan, J. Wu, and G. Shi, "Deep gaussian scale mixture prior for spectral compressive imaging," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 216–16 225.

[25] X. Hu, Y. Cai, J. Lin, H. Wang, X. Yuan, Y. Zhang, R. Timofte, and L. Van Gool, "Hdnet: High-resolution dual-domain learning for spectral compressive imaging," in *Proceedings of the IEEE/CVF*

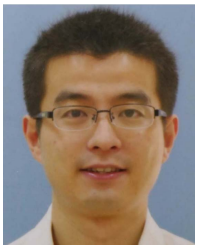
Conference on Computer Vision and Pattern Recognition, 2022, pp. 17 542–17 551.
[26] T. Klinghoffer, S. Somasundaram, K. Tiwary, and R. Raskar, "Physics vs. learned priors: Rethinking camera and algorithm design for task-specific imaging," in *2022 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2022, pp. 1–12.
[27] J. Bacca, T. Gelvez-Barrera, and H. Arguello, "Deep coded aperture design: An end-to-end approach for computational imaging tasks," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 1148–1160, 2021.
[28] X. Cao, H. Du, X. Tong, Q. Dai, and S. Lin, "A prism-mask system for multispectral video acquisition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 12, pp. 2423–2435, 2011.
[29] X. Lin, Y. Liu, J. Wu, and Q. Dai, "Spatial-spectral encoded compressive hyperspectral imaging," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 6, pp. 1–11, 2014.
[30] M. Descour and E. Dereniak, "Computed-tomography imaging spectrometer: experimental calibration and reconstruction results," *Applied optics*, vol. 34, no. 22, pp. 4817–4826, 1995.
[31] D. S. Jeon, S.-H. Baek, S. Yi, Q. Fu, X. Dun, W. Heidrich, and M. H. Kim, "Compact snapshot hyperspectral imaging with diffracted rotation," 2019.
[32] P.-J. Lapray, X. Wang, J.-B. Thomas, and P. Gouton, "Multispectral filter arrays: Recent advances and practical implementation," *Sensors*, vol. 14, no. 11, pp. 21 626–21 659, 2014.
[33] S. Mihoubi, O. Losson, B. Mathon, and L. Macaire, "Multispectral demosaicing using pseudo-panchromatic image," *IEEE Transactions on Computational Imaging*, vol. 3, no. 4, pp. 982–995, 2017.
[34] L. Wang, Z. Xiong, D. Gao, G. Shi, and F. Wu, "Dual-camera design for coded aperture snapshot spectral imaging," *Applied optics*, vol. 54, no. 4, pp. 848–858, 2015.
[35] L. Wang, Z. Xiong, D. Gao, G. Shi, W. Zeng, and F. Wu, "High-speed hyperspectral video acquisition with a dual-camera architecture," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4942–4950.
[36] X. Yuan, T.-H. Tsai, R. Zhu, P. Llull, D. Brady, and L. Carin, "Compressive hyperspectral imaging with side information," *IEEE Journal of selected topics in Signal Processing*, vol. 9, no. 6, pp. 964–976, 2015.
[37] Y. Fu, Y. Zheng, I. Sato, and Y. Sato, "Exploiting spectral-spatial correlation for coded hyperspectral image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3727–3736.
[38] Y. Liu, X. Yuan, J. Suo, D. J. Brady, and Q. Dai, "Rank minimization for snapshot compressive imaging," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 12, pp. 2990–3006, 2018.
[39] X. Yuan, "Generalized alternating projection based total variation minimization for compressive sensing," in *2016 IEEE International conference on image processing (ICIP)*. IEEE, 2016, pp. 2539–2543.
[40] L. Wang, C. Sun, Y. Fu, M. H. Kim, and H. Huang, "Hyperspectral image reconstruction using a deep spatial-spectral prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8032–8041.
[41] L. Wang, Z. Xiong, G. Shi, F. Wu, and W. Zeng, "Adaptive nonlocal sparse representation for dual-camera compressive hyperspectral imaging," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 10, pp. 2104–2111, 2016.
[42] H. Arguello, C. V. Correa, and G. R. Arce, "Fast lapped block reconstructions in compressive spectral imaging," *Applied Optics*, vol. 52, no. 10, pp. D32–D45, 2013.
[43] L. Wang, T. Zhang, Y. Fu, and H. Huang, "Hyperreconnet: Joint coded aperture optimization and image reconstruction for compressive hyperspectral imaging," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2257–2270, 2019.
[44] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 532–17 541.
[45] I. Choi, M. Kim, D. Gutierrez, D. Jeon, and G. Nam, "High-quality hyperspectral reconstruction using a spectral prior," Tech. Rep., 2017.
[46] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

Zhan Shi received the BS degree from the School of Computer Science and Engineering, Northeastern University, Liaoning, China, in 2021. He is currently working toward the MS degree with the School of Electronic Science and Engineering, Nanjing University, Nanjing, China. His research interests include computational photography and computer vision, especially computational spectral imaging.

Hao Ye received the BS degree from the School of Computer Science and Engineering, Northeastern University, Liaoning, China, in 2022. He is currently working toward the MS degree with the School of Electronic Science and Engineering, Nanjing University, Nanjing, China. His research interests include computational photography and computer vision.

Tao Lv received the BS degree from the college of Information Science and Engineering, Northeastern University, Liaoning, China, in 2021. He is currently working toward the Ph.D. degree with the School of Electronic Science and Engineering, Nanjing University, Nanjing, China. His research interests include computational photography and computer vision, especially computational spectral imaging.

Yibo Wang received the BS degree from the College of Electrical and Information Engineering, Hunan University, Changsha, China, in 2021. He is currently working toward the MS degree with the School of Electronic Science and Engineering, Nanjing University, Nanjing, China. His research interests include computer vision and deep learning.



Xun Cao (Member, IEEE) received the B.S. degree from Nanjing University, Nanjing, China, in 2006, and the Ph.D. degree from the Department of Automation, Tsinghua University, Beijing, China, in 2012. He held visiting positions with Philips Research, Aachen, Germany, in 2008, and Microsoft Research Asia, Beijing, from 2009 to 2010. He was a Visiting Scholar with The University of Texas at Austin, Austin, TX, USA, from 2010 to 2011. He is currently a Professor with the School of Electronic Science and Engineering, Nanjing University. His research interests include computational photography, image-based modeling and rendering, and VR/AR systems.