

Less Spectral Data, More Degradations: Rethinking Generalization Problem in CASSI Reconstruction

Xiaodong Wang^{1,2}, Yifan Si^{3,4†}, and Xin Yuan^{2†}

¹ Zhejiang University, Hangzhou, China

² Westlake University, Hangzhou, China

{wangxiaodong@westlake.edu.cn, xyuan@westlake.edu.cn}

³ School of Photonics and Optical Engineering,

Aerospace Information Technology University, Jinan, China

⁴ Shandong Key Laboratory of Intelligent Photonic Transmission
and Sensing, Jinan, China
siyifan@aitech.edu.cn

Abstract. Coded aperture snapshot spectral imaging (CASSI) reconstructs a hyperspectral cube from a single compressive measurement. Modern CASSI reconstruction networks have achieved impressive reconstruction quality, but they still suffer from a severe generalization problem: models trained on simulated measurements often perform poorly on real scenes or other unseen data distributions. A common belief is that the limited distribution of clean hyperspectral scenes is the main bottleneck, and that increasing spectral data should improve generalization. We argue that degradation diversity is more critical: existing CASSI reconstruction algorithms tend to overfit the degradation distribution induced by a fixed forward model, rather than learning a degradation-robust spectral prior. We therefore construct accurate degradation samples by decomposing the CASSI forward model into three stages, Modulation, Dispersion, and Integration, and explicitly modeling mask blur, chromatic dispersion, sensor noise, saturation, and 8-bit quantization for measurement synthesis. With this degradation-diverse training data, models trained on fewer spectral scenes achieve better robustness on unseen degradations and real measurements than models trained on more spectra with a single mask. The conclusion is verified on both deep unfolding and end-to-end reconstruction networks, suggesting that robust CASSI training should broaden the degradation distribution, not only the clean spectral distribution.

Keywords: CASSI · hyperspectral image reconstruction · snapshot compressive imaging · generalization problem · inverse problem

[†] Corresponding Author

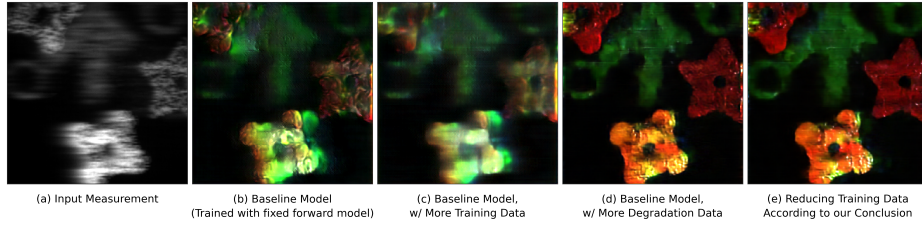


Fig. 1: CASSI generalization on real measurements. MST [2] is shown as an example. A baseline trained with a fixed forward model gives degraded real-data reconstruction. Simply increasing clean spectral training data does not remove the mismatch. Training with physically diverse degradations yields more reliable reconstruction, and reducing the spectral training set according to our conclusion remains effective.

1 Introduction

Coded aperture snapshot spectral imaging (CASSI) captures a three-dimensional hyperspectral image (HSI) with a two-dimensional sensor. Its forward process contains three physical operations: a coded aperture modulates the spectral cube, a dispersive element shifts different wavelengths, and the camera integrates the shifted coded bands into one compressive measurement [17]. Reconstructing the original HSI from this single measurement is a highly ill-posed inverse problem. Classical methods solve it with optimization and hand-crafted priors [6, 21]. Recent learning-based methods have become dominant, especially end-to-end networks and deep unfolding networks, which combine physical measurement models with learned priors and achieve strong benchmark performance [2–4, 10, 12, 13, 16, 18].

Despite their success, CASSI networks still suffer from a generalization problem: their reconstruction quality often drops when the test data distribution differs from the training data distribution, especially for real measurements. A common explanation is that hyperspectral training sets are small, e.g., CAVE has 241 scenes and KAIST has 30 scenes [5, 20]. Therefore, increasing the quantity and diversity of clean spectral scenes is often considered the direct way to improve generalization. We revisit this assumption by firstly analyzing how CASSI training data are generated. In a standard simulation protocol, clean spectral cubes pass through coded aperture modulation, uniform spectral dispersion, and sensor integration to form compressive measurements. This process involves two coupled distributions: the clean spectral data distribution and the degraded measurement distribution induced jointly by the spectra and the forward model. Since generalization measures reconstruction performance on unseen data distributions, it can be limited by either axis. In real systems, the degradation distribution is much richer than the ideal simulator: the mask may be blurred or noisy, dispersion may contain chromatic aberration, and the camera may introduce noise, saturation, and quantization. Thus, improving CASSI generalization requires studying not only clean spectral data diversity, but also degradation diversity.

We conduct comprehensive simulation and real-data experiments to study how clean spectral data and degradation diversity affect the generalization problem of CASSI. Figure 1 evaluates networks trained with a conventional fixed-degradation simulator directly on real measurements, without fine-tuning. Increasing the number of clean spectral scenes while using the same single degradation still gives poor results on an unseen physical degradation. In contrast, increasing degradation diversity significantly improves generalization. More interestingly, when the degradation set is sufficiently diverse, reducing the spectral training set to only 16 scenes still produces strong generalization. This may be because the baseline model is trained with a single fixed degradation, allowing the network to take an easier shortcut and overfit the simpler degradation distribution rather than the clean spectral distribution [7, 8]. Adding more clean spectra under the same fixed degradation does not remove this shortcut. In contrast, training with diverse degradations prevents the network from relying on one fixed measurement pattern and encourages it to learn more spectral-spatial priors. This does not imply that clean spectral diversity is unimportant. Rather, under forward-model mismatch, clean spectral diversity remains useful, but when the dominant test-time shift comes from the physical operator, covering the degradation distribution is more effective than simply increasing the number of clean spectral samples.

Our contributions are summarized as follows.

- (i) We study the generalization problem of CASSI reconstruction and find that existing methods tend to overfit the degradation distribution produced by a fixed forward model, rather than only suffering from insufficient clean spectral data.
- (ii) We construct a more accurate degradation data generation pipeline by decomposing the CASSI forward model into Modulation, Dispersion, and Integration, and modeling the non-idealities in each step.
- (iii) We conduct thorough experiments on four deep unfolding networks and two end-to-end architectures. The results on unseen synthetic degradations and real measurements show that, under forward-model mismatch, degradation diversity can be more important than simply increasing clean spectral data.

2 Related Work

2.1 CASSI Reconstruction

CASSI reconstruction has been studied from both model-based and learning-based perspectives. Early model-based methods formulate the inverse problem with sparse, low-rank, total-variation, or projection priors, and solve it by iterative optimization [11, 15, 21]. These methods are interpretable and can adapt to a calibrated forward model, but they are usually computationally expensive and depend strongly on the accuracy of the assumed operator. Learning-based approaches improve reconstruction speed and quality by training neural networks on simulated CASSI measurements. End-to-end models such as TSA-Net [13],

MST [2], and SPECAT [19] directly learn the inverse mapping from measurement to spectral cube. Deep unfolding methods such as GAP-Net [12], DAUHST [3], SSR [22], and MiJUN [14] embed the CASSI forward model into iterative reconstruction stages and learn the regularizer or update rule. These networks have become the strongest performers on benchmark simulations, but most of them are still trained with one fixed mask and one idealized simulator. Our work focuses on the generalization risk created by this narrow degradation distribution.

2.2 Generalization and Overfitting in Image Restoration

Generalization problems have been widely observed in image restoration, where a network trained on one synthetic degradation often fails under another degradation or on real data. This is closely related to shortcut learning: deep networks may exploit easy but non-causal correlations in the training distribution instead of learning the intended invariant structure [7]. Similar observations have been made in domain generalization studies [9]. In image deraining, Gu et al. [8] showed that networks can overfit the rain degradation distribution and generalize poorly when the degradation changes. CASSI has an analogous but more structured issue: the measurement is generated by a physical operator composed of mask coding, dispersion, and sensor integration. Therefore, the training distribution is determined not only by the clean spectral images, but also by the degradation/operator distribution. We bring this perspective to CASSI and show that enriching the degradation axis can be more important than simply collecting more clean spectra.

3 Method

3.1 Ideal CASSI Forward Model

The CASSI system acquires a two-dimensional compressive measurement by optically encoding and integrating a three-dimensional hyperspectral image. Let $\mathbf{X} \in \mathbb{R}^{H \times W \times B}$ denote the target hyperspectral cube, where H , W , and B are the spatial height, spatial width, and number of spectral bands, respectively. A physical binary coded aperture $\mathbf{M} \in \mathbb{R}^{H \times W}$ first modulates each spectral slice $\mathbf{X}_b \in \mathbb{R}^{H \times W}$ by element-wise multiplication:

$$\mathbf{X}'_b = \mathbf{M} \odot \mathbf{X}_b, \quad b = 1, \dots, B. \quad (1)$$

Due to the dispersion optics, each modulated spectral slice is laterally shifted by a wavelength-dependent offset d_b , resulting in

$$\mathbf{X}''(h, w, b) = \mathbf{X}'(h, w + d_b, b). \quad (2)$$

The sensor then integrates all shifted slices along the spectral dimension to produce the final two-dimensional measurement $\mathbf{Y} \in \mathbb{R}^{H \times (W + d_{\max})}$, where $d_{\max} =$

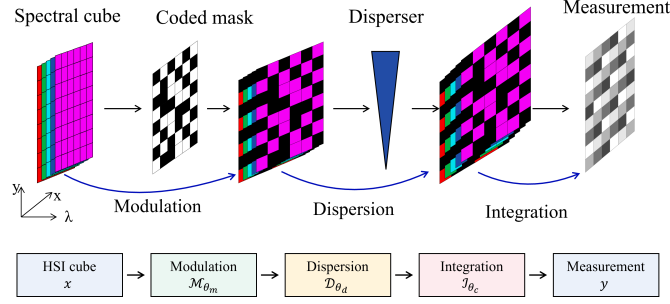


Fig. 2: CASSI forward model with three physical stages: modulation, dispersion, and integration.

$\max_b d_b$. In the presence of additive noise $\mathbf{N}$, the forward imaging process is

$$\mathbf{Y}(h, w) = \sum_{b=1}^B \mathbf{X}''(h, w, b) + \mathbf{N}(h, w). \quad (3)$$

After vectorization, this process can be compactly written as

$$\mathbf{y} = \Phi \mathbf{x} + \boldsymbol{\eta}, \quad \Phi = [S_{d_1} \text{Diag}(\mathbf{m}), S_{d_2} \text{Diag}(\mathbf{m}), \dots, S_{d_B} \text{Diag}(\mathbf{m})]. \quad (4)$$

Here $\mathbf{x} \in \mathbb{R}^{HWB}$ and $\mathbf{y} \in \mathbb{R}^{H(W+d_{\max})}$ are the vectorized forms of $\mathbf{X}$ and $\mathbf{Y}$, $\mathbf{m}$ is the vectorized form of $\mathbf{M}$, S_{d_b} is the matrix form of the shift operator, and $\Phi \in \mathbb{R}^{H(W+d_{\max}) \times HWB}$ is the system sensing matrix that encapsulates modulation, dispersion, and integration. Recovering $\mathbf{x}$ from $\mathbf{y}$ is a severely ill-posed inverse problem. Standard benchmark training fixes $\Phi = \Phi_0$ by using one binary mask, one deterministic shift rule, and noise-free summation. The key question in this paper is what happens when the test operator differs from this fixed Φ_0 .

3.2 Decomposing the Forward Operator

Equation (4) describes the ideal CASSI simulator as a linear sensing matrix Φ . This notation is convenient for reconstruction algorithms, but real measurement formation is more complicated: the effective mask can be blurred or noisy, the dispersion can contain wavelength-dependent PSFs, and the sensor can introduce saturation and quantization. We therefore use a parameterized forward operator $\mathcal{A}_\theta(\cdot)$ to represent practical degradation. As shown in Fig. 2, we decompose this operator according to the three physical operations that generate the measurement:

$$\mathbf{Y} = \mathcal{A}_\theta(\mathbf{X}), \quad \mathcal{A}_\theta = \mathcal{I}_{\theta_c} \circ \mathcal{D}_{\theta_d} \circ \mathcal{M}_{\theta_m}, \quad (5)$$

where $\mathcal{M}_{\theta_m}$ is mask modulation, $\mathcal{D}_{\theta_d}$ is chromatic dispersion, and $\mathcal{I}_{\theta_c}$ is sensor integration. The ideal benchmark simulator is a special case in which the mask is binary and sharp, dispersion is a deterministic shift, and integration is a noise-free summation. Our simulator enriches each stage with physically motivated

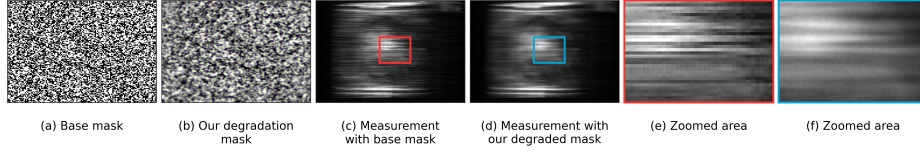


Fig. 3: Measurement synthesis from the binary mask under forward degradations: (a) the base mask, (b) its degraded mask, (c,d) measurements generated by the ideal and proposed degraded mask, and (e,f) corresponding zoomed regions. Modeling physical degradations produces measurements that better reflect real imaging conditions.

non-idealities. Fig. 3 compares the ideal measurement synthesis with our physical degradation pipeline. The proposed pipeline produces more realistic degraded masks and measurements, avoiding the overly regular periodic patterns caused by the conventional deterministic shift-and-sum simulator.

Modulation with spatially varying mask blur. The ideal modulation stage multiplies each band by a binary coded aperture. In practice, defocus, imperfect conjugacy, and fabrication errors make the effective mask spatially blurred, often sharper near the optical axis and blurrier toward the boundary. We model this stage as

$$\begin{aligned}\bar{\mathbf{M}}_b(u, v) &= \sum_{\Delta u, \Delta v} \mathbf{K}_m(u, v, \Delta u, \Delta v; \boldsymbol{\Sigma}_m(u, v)) \mathbf{M}_b(u - \Delta u, v - \Delta v), \\ \mathbf{Z}_b(u, v) &= \mathbf{X}_b(u, v) \bar{\mathbf{M}}_b(u, v), \\ \boldsymbol{\Sigma}_m(u, v) &= \sigma_{m,0} + (\sigma_{m,1} - \sigma_{m,0}) \mathbf{R}(u, v)^\rho.\end{aligned}\tag{6}$$

Here $\mathbf{K}_m$ is a normalized Gaussian PSF, $\mathbf{R}(u, v) \in [0, 1]$ is the normalized radial distance from the image center, and $\sigma_{m,0}, \sigma_{m,1}, \rho$ control the center blur, boundary blur, and radial growth rate. The variable modulation parameters are denoted by $\theta_m = \{\mathbf{M}_b, \mathbf{K}_m, \boldsymbol{\Sigma}_m, \sigma_{m,0}, \sigma_{m,1}, \rho\}$.

Dispersion with chromatic PSF. The ideal dispersion stage shifts band b by $s_b = s(b - 1)$. A prism-based system also introduces wavelength-dependent defocus and transverse chromatic aberration. We keep the nominal shift for comparability, but convolve each shifted band with a spectral-wise PSF:

$$\begin{aligned}\hat{\mathbf{Z}}_b(u, v) &= \sum_{\Delta u, \Delta v} \mathbf{K}_d(\lambda_b, \Delta u, \Delta v; \sigma_d(\lambda_b)) \mathbf{Z}_b(u - \Delta u, v - s_b - \Delta v), \\ \mathbf{K}_d(\lambda_b, \Delta u, \Delta v) &\propto \exp\left(-\frac{\Delta u^2}{2\sigma_{u,b}^2} - \frac{\Delta v^2}{2\sigma_{v,b}^2}\right), \\ \sigma_{u,b} &= \sigma_{u,0} + \alpha_u \left| \frac{\lambda_b - \lambda_0}{\lambda_{\max} - \lambda_{\min}} \right|^\gamma, \quad \sigma_{v,b} = \sigma_{v,0} + \alpha_v \left| \frac{\lambda_b - \lambda_0}{\lambda_{\max} - \lambda_{\min}} \right|^\gamma.\end{aligned}\tag{7}$$

This makes bands far from the reference focus blurrier than the central bands, while preserving the same nominal dispersion shift used in standard CASSI simulation. The variable dispersion parameters are $\theta_d = \{s_b, \mathbf{K}_d, \sigma_d, \sigma_{u,0}, \sigma_{v,0}, \alpha_u,$

$\alpha_v, \lambda_0, \gamma\}$. Here s_b is the nominal lateral shift of band b , $\mathbf{K}_d$ is the chromatic PSF, and $\sigma_d(\lambda_b)$ denotes its wavelength-dependent blur scale. The constants $\sigma_{u,0}$ and $\sigma_{v,0}$ define the base PSF width along the two spatial axes, α_u and α_v control how fast the blur grows away from the reference wavelength, λ_0 is the in-focus wavelength, and γ controls the chromatic growth curve.

Integration with camera noise and quantization. The sensor integrates all dispersed bands into one two-dimensional measurement and maps irradiance to digital counts:

$$\begin{aligned}\mathbf{E}(u, v) &= g \sum_{b=1}^B \hat{\mathbf{Z}}_b(u, v), \\ \mathbf{A}(u, v) &= \text{clip}(\mathbf{E}(u, v) + \mathbf{N}_{\text{shot}}(u, v) + \mathbf{N}_{\text{read}}(u, v), 0, C_{\text{max}}), \\ \mathbf{Y}(u, v) &= Q_L(\mathbf{A}(u, v)), \\ Q_L(a) &= \frac{\text{round}((2^L - 1)a/C_{\text{max}})}{2^L - 1}.\end{aligned}\tag{8}$$

Here g is exposure gain, $\mathbf{N}_{\text{shot}}$ is signal-dependent photon noise, $\mathbf{N}_{\text{read}} \sim \mathcal{N}(0, \sigma_r^2)$ is read noise, C_{max} is the saturation capacity, and Q_L is an L -bit quantizer. In our physical simulator, $L = 8$. The variable integration parameters are $\theta_c = \{g, \mathbf{N}_{\text{shot}}, \mathbf{N}_{\text{read}}, \sigma_r, C_{\text{max}}, L\}$.

3.3 Clean Spectral Data versus Degradation Data

Each CASSI training sample combines a clean spectral cube sampled from the clean-data distribution with a forward operator sampled from the degradation distribution. Let $P_{\mathbf{X}}$ denote the distribution of clean spectral cubes and $P_{\mathbf{A}}$ denote the distribution of CASSI forward operators. For a training pair, we first sample $\mathbf{X} \sim P_{\mathbf{X}}$ and $\mathbf{A} \sim P_{\mathbf{A}}$, then synthesize the measurement $\mathbf{Y} = \mathbf{A}(\mathbf{X})$. A reconstruction network f_ϕ takes the measurement and the corresponding operator as input, and is trained with a supervised spectral loss between the reconstructed HSI and the ground-truth HSI:

$$\min_{\phi} \mathbb{E}_{\mathbf{X} \sim P_{\mathbf{X}}, \mathbf{A} \sim P_{\mathbf{A}}} [\ell(f_\phi(\mathbf{A}(\mathbf{X}), \mathbf{A}), \mathbf{X})].\tag{9}$$

Here the loss is applied in the spectral image domain, i.e., to the final reconstructed HSI and the ground-truth HSI, which is the setting used in our experiments.

This formulation shows that CASSI training has two independent data axes. The first axis is the clean spectral data distribution $P_{\mathbf{X}}$, which is controlled by the number and diversity of clean HSI scenes. The second axis is the degradation distribution $P_{\mathbf{A}}$, which is controlled by the mask, blur, dispersion, sensor noise, saturation, and quantization used to generate measurements. Most benchmark protocols implicitly set $P_{\mathbf{A}} = \delta(\mathbf{A}_0)$, a point mass at one fixed operator. This works when training and testing share the same simulator, but it creates a mismatch when the test operator changes due to mask calibration, defocus,

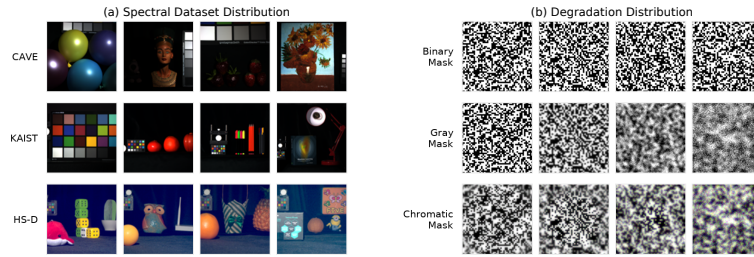


Fig. 4: CASSI training data contain two distributions: (a) clean spectral data and (b) mask-related degradation information.

dispersion, or sensor conditions. Therefore, increasing only the number of clean spectral scenes cannot cover a missing degradation distribution. Our experiments control these two axes separately to study which one matters more for generalization to unseen degradations.

3.4 Degradation-Diverse Training

We implement degradation-diverse training as operator-space augmentation by sampling $P_{\mathbf{A}}$ through modulation blur/noise, chromatic dispersion PSFs, and integration noise or quantization. This targets test-time operator shifts beyond clean-image perturbations. We evaluate this training on unseen simulated degradations and real measurements to investigate the generalization problem of CASSI.

4 Experiments

4.1 Experimental Setup

Base Mask and Degradation Settings We use one fixed binary mask as a shared base mask. What changes across protocols is not the mask pattern itself, but the effective degradation applied to the forward model. We use three degradation families. A Binary Mask is the ideal simulator: the base mask is sharp and binary, dispersion is a deterministic shift, and integration is a noiseless summation. A Gray Mask keeps the same base mask but samples achromatic Gaussian mask blur and Gaussian measurement noise during training. This family models simple mask defocus and sensor noise without wavelength-dependent effects. A Chromatic Mask further includes spatially varying mask blur and wavelength-dependent chromatic PSF; its full physical version also includes photon/read noise, saturation, and 8-bit quantization. Fig. 4(b) visualizes these degradation families, and Table 1(b) lists the corresponding degradation settings and parameter ranges. We use Gray Masks for degradation-diverse training in the main simulation study and reserve the Chromatic Mask as an unseen test degradation. For synthetic evaluation, the Chromatic Mask used for training/testing

Table 1: Training protocols and degradation families.

(a) Training protocols					(b) Degradation parameters		
Scenario	Abbrev.	Full name	Training dataset	Training degradation	Degradation	Forward model	Parameters
Simulation	Full-Binary	Full data + Binary Mask	256 scenes	Binary Mask	Binary Mask	Fixed binary mask, ideal shift, noiseless integration	None
	Full-Gray	Full data + Gray Mask	256 scenes	Gray Mask	Gray Mask	Gaussian mask blur and Gaussian measurement noise	$\sigma_m \sim \mathcal{U}(0.2, 2.4)$; $\sigma_n \sim \mathcal{U}(0.005, 0.12)$
	Few-Gray	Few data + Gray Mask	16 scenes	Gray Mask	Chromatic Mask	Spatially varying mask blur, chromatic PSF, photon/read noise, saturation, and 8-bit quantization	$\sigma_{m,0} = 0.12$; $\sigma_{m,1} = 3.0$; $\sigma_d(\lambda) \in [0.25, 2.8]$; $\sigma_r = 16e^-$; full-well = $8000e^-$; $g = 1.2$; $L = 8$
Real-data	Few-Chroma	Few data + Chromatic Mask	16 scenes	Chromatic Mask			

Table 2: Results using different training protocols on KAIST under the unseen Chromatic Mask degradation (PSNR $\uparrow$ /SSIM $\uparrow$ /SAM $\downarrow$).

Model	Train	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg.
GAP-Net	Full-Binary	24.40/0.70/0.37	21.97/0.59/0.44	18.60/0.60/0.51	25.66/0.73/0.57	19.67/0.55/0.42	20.90/0.62/0.54	21.02/0.61/0.45	21.39/0.66/0.53	19.76/0.62/0.58	21.62/0.57/0.50	21.50/0.63/0.49
	Full-Gray	27.19/0.80/0.36	23.57/0.67/0.44	21.64/0.77/0.39	29.46/0.85/0.40	22.19/0.73/0.39	26.36/0.79/0.45	23.12/0.71/0.42	24.56/0.78/0.43	24.25/0.76/0.38	24.68/0.71/0.40	24.70/0.76/0.41
	Few-Gray	28.10/0.81/0.33	23.66/0.67/0.42	22.84/0.80/0.36	31.11/0.87/0.36	22.18/0.73/0.41	26.44/0.79/0.46	23.76/0.72/0.39	24.58/0.78/0.44	24.85/0.77/0.36	25.13/0.72/0.40	25.27/0.77/0.39
DAUHST	Full-Binary	25.07/0.70/0.41	22.16/0.58/0.45	18.32/0.64/0.52	25.07/0.61/0.90	21.12/0.64/0.48	21.51/0.58/0.71	21.67/0.66/0.44	21.83/0.62/0.65	20.67/0.61/0.69	22.39/0.58/0.72	21.98/0.62/0.60
	Full-Gray	28.16/0.85/0.30	25.54/0.75/0.35	27.21/0.85/0.25	31.60/0.89/0.29	23.01/0.77/0.32	27.37/0.87/0.29	25.03/0.80/0.28	27.37/0.88/0.30	25.48/0.83/0.29	27.36/0.89/0.23	26.81/0.84/0.29
	Few-Gray	28.91/0.86/0.24	26.62/0.78/0.28	25.74/0.83/0.27	29.57/0.88/0.28	23.43/0.77/0.31	27.39/0.86/0.28	25.45/0.81/0.25	25.63/0.86/0.30	25.89/0.84/0.27	27.00/0.88/0.22	26.56/0.84/0.27
SSR	Full-Binary	24.37/0.69/0.36	22.45/0.60/0.41	16.96/0.55/0.59	24.07/0.67/0.71	20.68/0.59/0.39	21.72/0.62/0.59	20.69/0.58/0.48	21.32/0.63/0.59	19.49/0.59/0.58	19.94/0.59/0.65	21.19/0.61/0.54
	Full-Gray	29.90/0.89/0.20	27.21/0.82/0.29	30.98/0.91/0.17	38.41/0.97/0.16	25.77/0.83/0.28	28.69/0.89/0.28	27.91/0.87/0.19	27.02/0.87/0.34	28.21/0.88/0.22	30.04/0.92/0.23	29.41/0.88/0.23
	Few-Gray	29.60/0.88/0.21	27.45/0.80/0.39	29.40/0.89/0.20	35.43/0.95/0.17	25.64/0.84/0.28	29.10/0.89/0.28	26.76/0.86/0.22	27.97/0.88/0.36	29.19/0.89/0.20	29.67/0.92/0.24	29.02/0.88/0.26
MiJUN	Full-Binary	26.15/0.79/0.30	23.08/0.66/0.38	20.01/0.72/0.45	25.80/0.81/0.45	22.17/0.73/0.30	22.35/0.73/0.41	22.28/0.69/0.41	21.43/0.74/0.41	21.00/0.73/0.42	22.97/0.71/0.41	22.73/0.73/0.39
	Full-Gray	30.61/0.89/0.23	28.14/0.82/0.26	28.28/0.89/0.20	36.22/0.94/0.22	26.86/0.86/0.20	30.81/0.91/0.23	28.08/0.86/0.19	28.94/0.89/0.28	29.28/0.89/0.20	28.82/0.91/0.20	29.61/0.89/0.22
	Few-Gray	31.17/0.90/0.21	29.10/0.83/0.23	30.51/0.91/0.16	35.40/0.94/0.20	26.47/0.86/0.21	30.95/0.92/0.21	30.13/0.89/0.15	28.41/0.90/0.27	30.73/0.90/0.18	28.75/0.90/0.20	30.16/0.90/0.20
MST	Full-Binary	22.70/0.69/0.58	20.63/0.57/0.66	17.58/0.61/0.70	23.64/0.76/0.62	18.22/0.61/0.55	20.08/0.71/0.47	18.81/0.56/0.75	20.98/0.72/0.52	18.25/0.67/0.64	20.97/0.70/0.47	20.19/0.66/0.60
	Full-Gray	27.08/0.83/0.37	25.88/0.78/0.32	25.50/0.83/0.31	31.90/0.90/0.29	21.27/0.74/0.44	26.71/0.87/0.27	23.94/0.78/0.33	27.26/0.88/0.29	24.38/0.82/0.34	25.61/0.87/0.25	25.95/0.83/0.32
	Few-Gray	27.08/0.84/0.34	25.65/0.77/0.32	24.85/0.83/0.29	32.42/0.91/0.31	21.35/0.74/0.44	27.56/0.88/0.27	24.09/0.78/0.31	26.78/0.87/0.32	24.56/0.82/0.34	24.88/0.86/0.24	25.92/0.83/0.32
SPECAT	Full-Binary	22.23/0.62/0.65	20.43/0.54/0.74	16.90/0.58/0.80	24.77/0.74/0.80	17.23/0.55/0.64	19.33/0.66/0.59	18.66/0.54/0.82	19.74/0.68/0.60	17.52/0.62/0.73	19.44/0.60/0.61	19.62/0.61/0.70
	Full-Gray	28.02/0.85/0.28	26.33/0.76/0.31	25.57/0.83/0.26	33.86/0.90/0.28	24.04/0.80/0.28	28.71/0.88/0.27	25.29/0.80/0.25	27.66/0.87/0.35	25.21/0.82/0.31	26.74/0.87/0.25	27.14/0.84/0.28
	Few-Gray	28.43/0.85/0.29	26.32/0.77/0.29	26.16/0.85/0.25	34.13/0.91/0.24	24.18/0.79/0.29	29.35/0.88/0.25	25.54/0.81/0.25	27.92/0.87/0.30	25.55/0.84/0.27	26.83/0.86/0.26	27.44/0.84/0.27

variants is optical-only and does not include camera degradation; for real-data evaluation, the measured data naturally include the full camera degradation.

Training Allocations We use three training protocols, named by clean-data amount and mask family. **Full-Binary** uses the full clean spectral set with one Binary Mask, following conventional fixed-degradation training. Our protocols with diverse degradations, **Full-Gray** and **Few-Gray**, use the same random Gray Mask distribution; the former uses the full clean spectral set, and the latter uses only 16 scenes. All models are trained for 300 epochs using the same reconstruction supervision on the final output stage. For deep unfolding networks, the corresponding mask is provided to the network in the same format as the forward model. Table 1(a) summarizes the abbreviations. The degradation ranges are chosen to cover mild-to-severe practical operator mismatch and are not tuned on the KAIST test scenes or real measurements.

Architectures & Metrics We test both deep unfolding and end-to-end CASSI reconstruction networks: GAPNet [12], DAUHST [3], SSR [22], MiJUN [14], MST [2], and SPECAT [19]. The four deep unfolding models cover different prior modules: GAPNet is a CNN-based unfolding baseline, DAUHST uses spatial attention, SSR emphasizes spectral attention, and MiJUN uses a Mamba-inspired sequence module. MST and SPECAT are end-to-end transformer-based

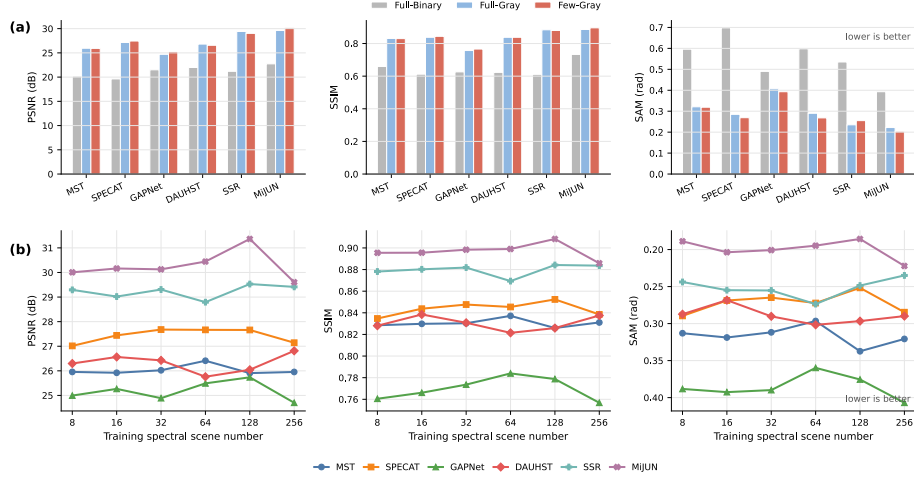


Fig. 5: Generalization analysis: (a) degradation-diverse training (Full-Gray & Few-Gray) improves reconstruction under forward-model mismatch; (b) increasing clean spectral scenes brings only mild changes across unseen degradation tests.

reconstruction networks. Since different architectures require different input formats, we follow the input convention of each original method when feeding measurements and masks. We report peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and spectral angle mapper (SAM) to evaluate reconstruction fidelity, structural consistency, and spectral accuracy, respectively.

Implementation Details The clean training pool is constructed from CAVE, KAIST, and about 18 HS-D scenes [1], resulting in 256 spectral scenes in total. All models are trained on 256×256 crops with 28 spectral bands and evaluated on the 10 KAIST test scenes. Unless otherwise specified, training uses Adam with batch size 5, 5000 random crops per epoch, 300 epochs, initial learning rate 4×10^{-4} , and cosine annealing to 10^{-6} . The loss is the root mean squared error between the final reconstruction and the ground-truth HSI. Full-Binary uses all 256 scenes with the fixed Binary Mask. Full-Gray uses the same 256 scenes but samples Gray Mask blur and noise online for each training batch. Few-Gray uses 16 scenes with the same Gray Mask sampling range. All synthetic generalization results are evaluated on the held-out Chromatic Mask degradation defined in Table 1(b); this degradation is not used by Full-Binary, Full-Gray, or Few-Gray during training.

Definition of Generalization We measure generalization by PSNR, SSIM, and SAM on the unseen Chromatic Mask degradation in Table 1(b). Better reconstruction under this held-out operator indicates stronger generalization.

4.2 Quantitative and Qualitative Evaluation

Simulation Results on Generalization We first evaluate whether the training protocols in Table 1(a) transfer to an unseen Chromatic Mask degrada-

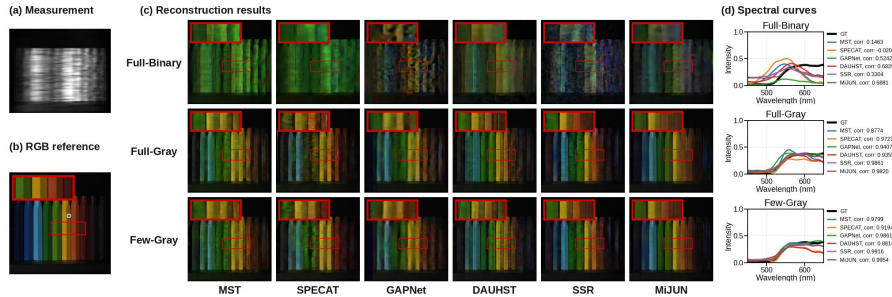


Fig. 6: Simulation results on generalization with scene 09: (a) measurement, (b) RGB reference, (c) reconstructions, and (d) ROI spectra. Degradation-diverse training (Full-Gray & Few-Gray) improves spatial details and spectral accuracy.

tion on KAIST. The Chromatic Mask is not used during Gray Mask training and thus can be used as a generalization test. Table 2 reports scene-wise PSNR/SSIM/SAM for six reconstruction architectures. The trend is consistent across both end-to-end and deep unfolding models: Full-Binary, which sees many clean spectra but only one ideal Binary Mask, suffers large performance degradation on the Chromatic Mask. Full-Gray improves the average spatial and spectral metrics by exposing the model to random achromatic degradations during training. Few-Gray, despite using only a small clean spectral subset, remains competitive with Full-Gray and is consistently stronger than Full-Binary under the unseen operator.

The previous results show that changing the degradation distribution is effective. We next ask whether simply increasing the number of clean spectral scenes can provide the same generalization. Fig. 5 answers this question from two views. In Fig. 5(a), the grouped bars summarize the mean KAIST metrics in Table 2. Adding Gray Mask degradations improves generalization from Full-Binary to Full-Gray, and reducing the clean spectral set from Full-Gray to Few-Gray preserves most of the gain. This indicates that degradation diversity is more influential than clean-data quantity for the unseen Chromatic Mask. Fig. 5(b) further varies the number of clean spectral scenes under the same random Gray Mask protocol, from 8 scenes to more than 200. The performance on the unseen degradation changes only mildly over this wide range, with small fluctuations across the six architectures. This suggests that, in the generalization test, the number of clean spectral training scenes is not the dominant factor once degradation diversity is available.

Fig. 6 visualizes the same conclusion on scene09. The measurement and RGB reference are shown on the left, while the reconstruction grid compares MST, SPECAT, GAPNet, DAUHST, SSR, and MiJUN under the three training protocols. Full-Binary often produces unstable colors and artifacts near fine text regions. Full-Gray and Few-Gray preserve sharper spatial structures and produce spectra that are more correlated with the ground truth in the selected region of interest (ROI). The important point is not that fewer clean scenes are

Table 3: Real-measurement proxy evaluation with reprojection error $E_{\text{repr}} \downarrow$.

Model	Training	Scene1	Scene2	Scene3	Scene4	Scene5	Avg.
MST	Full-Binary	0.168	0.241	0.212	0.190	0.199	0.202
	Few-Gray	0.124	0.211	0.191	0.160	0.148	0.167
	Few-Chroma	0.122	0.263	0.189	0.172	0.148	0.179
SPECAT	Full-Binary	0.187	0.250	0.235	0.198	0.192	0.212
	Few-Gray	0.180	0.250	0.222	0.181	0.173	0.201
	Few-Chroma	0.196	0.288	0.261	0.261	0.213	0.244
DAUHST	Full-Binary	0.205	0.550	0.600	0.379	0.322	0.411
	Few-Gray	0.108	0.185	0.172	0.133	0.126	0.145
	Few-Chroma	0.134	0.521	0.397	0.359	0.205	0.323

always better, but that degradation diversity removes a shortcut: the network can no longer rely on measurement statistics specific to one fixed mask.

Real-data Results Simulation experiments provide ground-truth PSNR, SSIM, and SAM, but real CASSI measurements reveal whether the trend transfers to practical acquisition. We therefore test the trained checkpoints on TSA real measurements with the calibrated real mask and the same preprocessing convention as common MST/DAUHST simulation protocols. Because real spectral ground truth is unavailable, we report both a weak physical-consistency proxy and qualitative reconstructions.

We compute the proxy by reprojecting each reconstruction through the calibrated CASSI forward model:

$$E_{\text{repr}} = \frac{\|\alpha^* A_{\text{calib}}(\hat{\mathbf{X}}) - \mathbf{Y}\|_1}{\|\mathbf{Y}\|_1}, \quad \alpha^* = \frac{\langle A_{\text{calib}}(\hat{\mathbf{X}}), \mathbf{Y} \rangle}{\|A_{\text{calib}}(\hat{\mathbf{X}})\|_2^2}. \quad (10)$$

Here α^* is a scalar exposure calibration for each real measurement. A lower E_{repr} indicates that the reconstructed cube better explains the captured measurement under the calibrated operator. Since the calibrated operator is still approximate, we use this metric only as proxy evidence.

Fig. 7 shows scene01, scene03, and scene05. For MST, SPECAT, and SSR, the fixed Full-Binary baseline gives conspicuous structured artifacts and unstable colors. Few-Gray is more stable, and **Few-Chroma**, trained with the fuller physical degradation family, further improves visual consistency. This real-data behavior agrees with the synthetic Chromatic Mask results: the failure mode is primarily a forward-operator mismatch rather than merely a lack of clean spectral scenes.

4.3 Ablation Study

Generalization across Various Degradations Table 4 studies how trained models behave under test degradations of different severity. We choose SSR as a representative deep unfolding model and SPECAT as a representative end-to-end model, and test each trained checkpoint under three component-wise test settings: Binary, Binary with mask blur, and Binary with both mask blur and chromatic PSF.

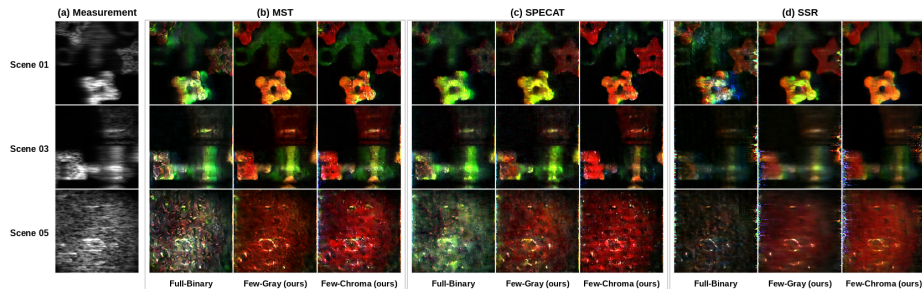


Fig. 7: Real-measurement reconstruction: (a) measurements, (b) MST, (c) SPECAT, and (d) SSR. Few-Gray and Few-Chroma improve real-data stability.

Table 4: Component-wise ablation of degradation modeling (PSNR $\uparrow$ /SSIM $\uparrow$ /SAM $\downarrow$).

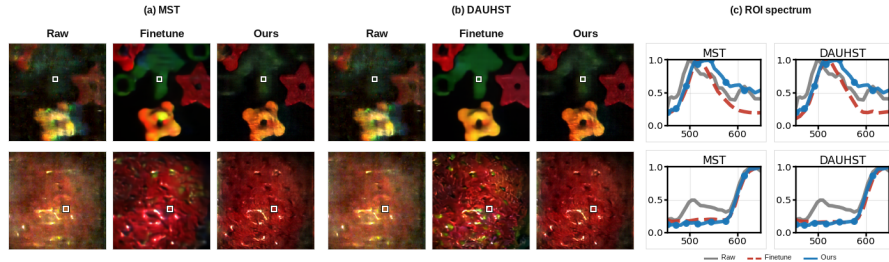
Model	Train	Binary	Binary + mask blur	Binary + mask blur + chromatic PSF
SPECAT	Full-Binary	34.16/0.934/0.168	19.73/0.615/0.688	19.62/0.612/0.698
	Full-Gray	32.12/0.895/0.220	30.86/0.898/0.199	27.14/0.838/0.285
	Few-Gray	31.85/0.892/0.216	30.75/0.898/0.184	27.44/0.844/0.269
SSR	Full-Binary	38.00/0.969/0.081	21.42/0.568/0.597	21.19/0.611/0.535
	Full-Gray	36.50/0.960/0.094	35.31/0.951/0.120	29.41/0.884/0.235
	Few-Gray	35.99/0.957/0.098	35.12/0.949/0.118	29.02/0.880/0.255

The table supports two conclusions. First, reconstruction becomes harder as the test degradation becomes stronger. Binary corresponds to the ideal simulator, the second setting adds mask blur, and the third further adds chromatic PSF. Thus the metric drop across the three columns reflects increasing forward-operator difficulty. Second, degradation-diverse training makes the model less sensitive to this change. For example, SPECAT trained with Full-Binary reaches 34.16 dB on Binary but drops to 19.62 dB when mask blur and chromatic PSF are both present. In contrast, SPECAT trained with Gray Mask degradations remains around 27 dB under this strongest setting, for both Full-Gray and Few-Gray. The same trend appears for SSR: the Binary-trained model has the best in-distribution Binary result, but Gray-trained models degrade much more gracefully when the test operator changes.

Real-data Fine-tuning Comparison Fig. 8 compares degradation-diverse training with fine-tuning on a calibrated real mask. The proposed model is directly evaluated on real measurements without target-scene fine-tuning. Table 5 reports the scene-wise reprojection error for the same comparison. Here Raw denotes the fixed-forward-model reconstruction, Finetune denotes five-epoch adaptation to the calibrated real mask, and Ours denotes Few-Gray training without any real-measurement fine-tuning. Therefore, Raw and Ours correspond to the Full-Binary and Few-Gray protocols in Table 3, respectively. Across the five real scenes, Ours gives lower average reprojection error for both MST and DAUHST, which is consistent with the visual comparison in Fig. 8.

Table 5: Real-data fine-tuning proxy comparison with $E_{\text{repr}} \downarrow$.

Model	Method	Scene1	Scene2	Scene3	Scene4	Scene5	Avg.
MST	Raw	0.168	0.241	0.212	0.190	0.199	0.202
	Finetune	0.158	0.380	0.293	0.273	0.228	0.266
	Ours	0.124	0.211	0.191	0.160	0.148	0.167
DAUHST	Raw	0.205	0.550	0.600	0.379	0.322	0.411
	Finetune	0.177	0.356	0.241	0.238	0.181	0.239
	Ours	0.108	0.185	0.172	0.133	0.126	0.145

**Fig. 8:** Real-data fine-tuning ablation with MST and DAUHST. Ours denotes Few-Gray without target-scene fine-tuning.

5 Conclusion

We studied the generalization problem of CASSI reconstruction from the perspectives of training data and forward modeling. Our analysis suggests that existing reconstruction networks can overfit a single degradation distribution induced by a fixed forward model, limiting generalization even when more clean spectral data are available. To address this mismatch, we decomposed the forward model into modulation, dispersion, and integration and modeled physical non-idealities at each stage to generate diverse training degradations. Experiments across multiple architectures show improved reconstruction under unseen simulated degradations and real measurements without fine-tuning, even with only 16 clean spectral scenes. Clean spectral diversity remains useful, but under forward-model mismatch, degradation diversity can be more influential than data quantity. These findings motivate designing CASSI training data jointly around clean spectral content and realistic forward degradations.

Acknowledgments. This work was supported by “Pioneer” and “Leading Goose” R&D Program of Zhejiang (grant 2024SDXHDX0006, 2024C03182). This work was supported by National Key R&D Program of China (2024YFF0505603) and the National Natural Science Foundation of China (grant number U2541205, 62271414). This work was supported by the Shandong Provincial Natural Science Foundation (Grant No. ZR2026QC2221Z).

References

1. Baek, S.H., Ikoma, H., Jeon, D.S., Li, Y., Heidrich, W., Wetzstein, G., Kim, M.H.: Single-shot hyperspectral-depth imaging with learned diffractive optics. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2651–2660 (2021)
2. Cai, Y., Lin, J., Hu, X., Wang, H., Yuan, X., Zhang, Y., Timofte, R., Van Gool, L.: Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17502–17511 (2022)
3. Cai, Y., Lin, J., Wang, H., Yuan, X., Ding, H., Zhang, Y., Timofte, R., Van Gool, L.: Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging. In: Advances in Neural Information Processing Systems (2022)
4. Chen, Y., Wang, Y., Wang, X., Yuan, X., Zhang, H.: LCTC: Lightweight convolutional thresholding sparse coding network prior for compressive hyperspectral imaging. *IEEE Transactions on Image Processing* **34**, 4286–4301 (2025). <https://doi.org/10.1109/TIP.2025.3583951>
5. Choi, I., Jeon, D.S., Nam, G., Gutierrez, D., Kim, M.H.: High-quality hyperspectral reconstruction using a spectral prior. *ACM Transactions on Graphics* **36**(6), 1–13 (2017)
6. Duarte, M.F., Davenport, M.A., Takhar, D., Laska, J.N., Sun, T., Kelly, K.F., Baraniuk, R.G.: Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine* **25**(2), 83–91 (2008)
7. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2**, 665–673 (2020)
8. Gu, J., Ma, X., Kong, X., Qiao, Y., Dong, C.: Networks are slacking off: Understanding generalization problem in image deraining. In: Advances in Neural Information Processing Systems. vol. 36, pp. 28565–28584 (2023)
9. Gulrajani, I., Lopez-Paz, D.: In search of lost domain generalization. In: International Conference on Learning Representations (2021)
10. He, Z., Wang, P., Wang, X., Chen, C., Yuan, X.: Joint spectral image reconstruction and semantic segmentation with cooperative unfolding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6910–6919 (2026)
11. Liu, Y., Yuan, X., Suo, J., Brady, D.J., Dai, Q.: Rank minimization for snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**(12), 2990–3006 (2019)
12. Meng, Z., Jalali, S., Yuan, X.: Gap-net for snapshot compressive imaging. In: European Conference on Computer Vision. pp. 108–124. Springer (2020)
13. Meng, Z., Ma, J., Yuan, X.: End-to-end low cost compressive spectral imaging with spatial-spectral self-attention. In: European Conference on Computer Vision. pp. 187–204. Springer (2020)
14. Qin, M., Feng, Y., Wu, Z., Zhang, Y., Yuan, X.: Detail matters: Mamba-inspired joint unfolding network for snapshot spectral compressive imaging. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 6594–6602 (2025)
15. Qiu, W., Meng, Z., Yuan, X.: Effective snapshot compressive-spectral imaging via deep denoising and total variation priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9127–9136 (2021)

16. Si, Y., Lin, Z., Wang, X., He, S.: A new hyperspectral reconstruction method with conditional diffusion model for snapshot spectral compressive imaging. *IEEE Transactions on Instrumentation and Measurement* **74**, 1–14 (2025). <https://doi.org/10.1109/TIM.2025.3551465>, article 4506214
17. Wagadarikar, A., John, R., Willett, R., Brady, D.: Single disperser design for coded aperture snapshot spectral imaging. *Applied Optics* **47**(10), B44–B51 (2008)
18. Wang, X., He, Z., Wang, P., Wang, L., Hu, Y., Yuan, X.: Spectral compressive imaging via chromaticity-intensity decomposition. In: *Advances in Neural Information Processing Systems*. vol. 38, pp. 4024–4046 (2025). <https://doi.org/10.52202/085713-0142>
19. Yao, Z., Liu, S., Yuan, X., Fang, L.: SPECAT: SPatial-spEctral Cumulative-Attention Transformer for high-resolution hyperspectral image reconstruction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 25368–25378 (2024)
20. Yasuma, F., Mitsunaga, T., Iso, D., Nayar, S.K.: Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum. *IEEE Transactions on Image Processing* **19**(9), 2241–2253 (2010)
21. Yuan, X.: Generalized alternating projection based total variation minimization for compressive sensing. *IEEE International Conference on Image Processing* pp. 2539–2543 (2016)
22. Zhang, J., et al.: Spatial-spectral reconstruction network for hyperspectral image compressive sensing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024)