DECOUPLED DEEP UNROLLING MODEL FOR SNAPSHOT SPECTRAL IMAGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Coded Aperture Snapshot Spectral Imaging (CASSI) compresses a 3D hyperspectral image (HSI) into a 2D measurement through cascaded mask modulation, spectral dispersion, and spectral integration. Existing deep unrolling methods typically represent these operations using a single coupled forward model and solve the resulting highly ill-posed reconstruction problem without dedicated regularization of intermediate physical states. To address these limitations, we propose the Decoupled Deep Unrolling Model (DDUM), which explicitly introduces intermediate state variables along the physical imaging chain of CASSI. Under a Bayesian formulation, DDUM decomposes reconstruction into three physically interpretable subproblems: integration inversion, spectral alignment, and mask demodulation. This decoupling enables dedicated regularization networks to model the distinct priors of different intermediate states, yielding more faithful reconstruction. To address the ambiguity inherent in many-to-one spectral integration, we further develop a learnable generalized physical inverse operator. We further analyze the inverse properties of the coupled and constituent operators. Extensive experiments demonstrate that DDUM consistently outperforms state-of-the-art methods and that adapting existing unrolling networks to our framework further improves their reconstruction performance.

1 INTRODUCTION

Hyperspectral imaging (HSI) captures rich spectral information beyond RGB images and has been widely used in many fields Du et al. (2016); Wang et al. (2020b); Maggiori et al. (2017). To accelerate data acquisition, various snapshot compressive imaging systems have been developed, among which Coded Aperture Snapshot Spectral Imaging (CASSI) has emerged as a promising solution.

As illustrated in fig. 1a, the CASSI process involves three cascaded processes: mask modulation, spectral dispersion, and spectral integration. While this imaging mechanism enables efficient snapshot acquisition, it progressively degrades and mixes the spatial–spectral information, making the recovery of a full 3D HSI cube from a 2D measurement highly challenging.

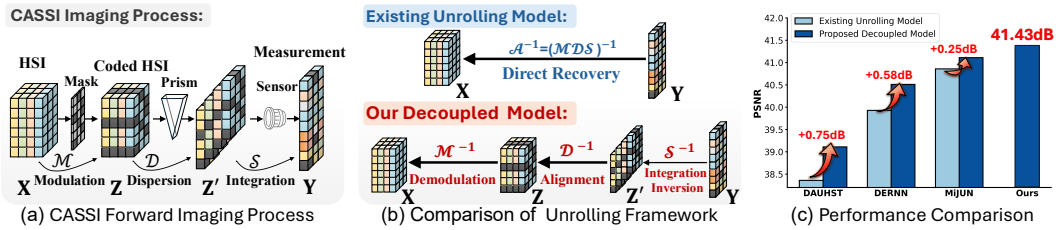
Various methods Yuan (2016); Miao et al. (2019) have been proposed for CASSI reconstruction, among which Deep Unrolling Models (DUMs) have shown particularly strong performance. Most existing DUMs collapse the cascaded CASSI process into a single coupled forward operator and formulate reconstruction as a regularized inverse problem:

$$\arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - \mathcal{A}\mathbf{x}\|_2^2 + \tau R(\mathbf{x}), \tag{1}$$

where $\mathbf{y}$ denotes the 2D measurement, $\mathbf{x}$ is the target 3D HSI, and τ balances data fidelity and regularization. The coupled operator $\mathcal{A} = \mathcal{S}\mathcal{D}\mathcal{M}$ comprises mask modulation $\mathcal{M}$, spectral dispersion $\mathcal{D}$, and spectral integration $\mathcal{S}$. The regularizer $R(\mathbf{x})$ encodes the HSI prior and is typically implemented using a learnable network.

Many existing DUMs Wang et al. (2020a); Zhang et al. (2024a) develop increasingly expressive regularization networks to capture HSI priors while retaining the coupled reconstruction formulation. Despite their effectiveness, such formulations have two limitations. First, existing coupled formulations jointly address multiple imaging degradations within a single, highly ill-posed

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) CASSI Forward Imaging Process (b) Comparison of Unrolling Framework (c) Performance Comparison

Figure 1: Overview of the proposed decoupled unrolling paradigm. (a) The CASSI forward process consists of three cascaded physical operations: mask modulation, spectral dispersion, and spectral integration. (b) Existing coupled formulations represent these operations using a single operator $\mathcal{A} = \mathcal{SDM}$ and reconstruct the target HSI without dedicated regularization of intermediate physical states. In contrast, DDUM explicitly models these states and decomposes reconstruction into three physically interpretable inverse subproblems. (c) DDUM improves reconstruction quality, and adapting existing unrolling networks to our framework yields further performance gains.

inverse problem. As shown in table 1, the coupled operator $\mathcal{A}$ exhibits a higher-dimensional null space and poorer conditioning than its constituent operators, making reliable recovery of the target HSI more challenging. Second, existing coupled formulations regularize only the target HSI $\mathbf{x}$, leaving the intermediate physical states after mask modulation ($\mathbf{z}$) and spectral dispersion ($\mathbf{z}'$) without dedicated regularization (fig. 1a). Consequently, the distinct structural priors of these states are not explicitly exploited to guide reconstruction, which can limit the fidelity of the recovered HSI.

Table 1: **CASSI Operator characteristics.**[1]

Op.	IRR	Null	κ	Cost
$\mathcal{M}$	p	$(1-p)N_\lambda n$	1	$\mathcal{O}(N_\lambda n)$
$\mathcal{D}$	1	0	1	-
$\mathcal{S}$	$1/N_\lambda$	$(N_\lambda-1)n$	1	$\mathcal{O}(N_\lambda n)$
$\mathcal{A}$	p/N_λ	$>(N_\lambda-1)n$	$\sqrt{N_\lambda}$	$\mathcal{O}((N_\lambda n)^3)$

To address these limitations, we propose the Decoupled Deep Unrolling Model (DDUM), a physically decoupled paradigm for CASSI reconstruction. Under a Bayesian formulation, DDUM explicitly introduces intermediate physical states and decomposes the coupled reconstruction problem into three physically interpretable subproblems: integration inversion, spectral alignment, and mask demodulation (see fig. 1b). Each intermediate state is regularized by a dedicated learnable prior, allowing its distinct structural characteristics to guide reconstruction. To address the ambiguity inherent in many-to-one spectral integration, we further develop a learnable generalized physical inverse operator. Our theoretical analysis characterizes the advantages of the decoupling framework from the perspective of operator properties. Extensive experiments demonstrate that DDUM consistently outperforms state-of-the-art methods in reconstruction fidelity. Moreover, adapting existing unrolling networks to the proposed framework improves their reconstruction performance, supporting its applicability across different architectures. Our contributions are summarized as follows:

1. We propose DDUM, a physically decoupled unrolling framework that explicitly models intermediate imaging states and decomposes CASSI reconstruction into three interpretable subproblems. We further provide a theoretical analysis of the operator properties underlying this decomposition.

2. We derive a Bayesian formulation with dedicated learnable priors for intermediate states and develop a learnable generalized physical inverse operator to address the ambiguity of many-to-one spectral integration.

3. Extensive experiments demonstrate that DDUM consistently outperforms state-of-the-art reconstruction methods. Adapting existing unrolling networks to the proposed framework also yields performance gains, supporting its applicability across different architectures.

[1] p: mask retention ratio; $n{=}H{\times}W$; N_λ: spectral channels. IRR: information retention ratio; Null: null-space dimension; κ: observable-subspace condition number. Cost denotes inverse mapping cost under the regularized normal equation; derivations are in the appendix.

2 Related Work

2.1 Deep Unrolling for CASSI Reconstruction

Since the pioneering model DNU Wang et al. (2020a), Deep Unrolling Model have become a dominant paradigm for CASSI reconstruction. Most existing DUMs formulate CASSI reconstruction as a coupled inverse problem and primarily improve reconstruction by developing increasingly powerful HSI priors, ranging from CNN-based networks Wang et al. (2019); Huang et al. (2021); Yuan et al. (2020) to Transformer-based architectures Cai et al. (2022b); Yao et al. (2024); Cai et al. (2022c;a); Zhang et al. (2024a) and, more recently diffusion models Wu et al. (2025). Despite these advances, the underlying coupled reconstruction formulation remains largely unchanged.

2.2 Decoupling and Physics-Guided Reconstruction

Decoupling strategies have been explored to separate restoration tasks or exploit structured image representations. For example, IDD-BM3D separates deblurring from denoising Danielyan et al. (2011), while IDDNet develops a learned iterative decoupling framework for image restoration Ji et al. (2024). In spectral compressive imaging, LRDUN integrates a low-rank HSI representation into the sensing model and reconstructs spectral bases and subspace images Huang et al. (2026). CIDNet instead decomposes the HSI into chromaticity and intensity components Wang et al. (2025). These approaches organize reconstruction around restoration tasks or image representations, rather than explicitly separating the inverse tasks associated with the cascaded CASSI imaging operations.

Physical information has also been incorporated into CASSI reconstruction through mask-guided attention in MST Cai et al. (2022b) and spectral–spatial rectification in SSR Zhang et al. (2024b). DDUM takes a different approach by explicitly decomposing the inverse problem along the CASSI imaging chain. By introducing variables for the physical states after mask modulation and spectral dispersion, DDUM formulates integration inversion, spectral alignment, and mask demodulation as distinct subproblems and assigns dedicated learnable priors to the intermediate states.

3 Method

In this section, we first characterize the CASSI imaging model and its optimization properties (Sec. 3.1), based on which we formulate a Bayesian decoupled reconstruction model with explicit intermediate states (Sec. 3.2) and develop an unrolling-based solver for its optimization (Sec. 3.3).

3.1 CASSI Forward Model and Operator Properties

Coded Aperture Snapshot Spectral Imaging records a 3D HSI cube $\mathbf{X} \in \mathbb{R}^{H\times W\times N_\lambda}$ (height H, width W, spectral channels N_λ) as a single 2D measurement through three cascaded physical operations:

(1) Mask modulation ($\mathcal{M}$). The mask modulation operator $\mathcal{M} = \{\mathbf{M}_{n_\lambda}\}_{n_\lambda=1}^{N_\lambda}$ consists of N_λ band-wise coded masks. Each $\mathbf{M}_{n_\lambda}$ is a matrix in $\mathbb{R}^{H\times W}$ with a retention ratio p, i.e., the proportion of non-zero elements is p. The operator $\mathcal{M}$ acts on each spectral band as follows:

$$\mathbf{Z}_{n_\lambda} = \mathbf{M}_{n_\lambda} \odot \mathbf{X}_{n_\lambda} + \mathbf{N}_{1,n_\lambda}, \quad n_\lambda = 1, \ldots, N_\lambda, \tag{2}$$

where $\mathbf{X}_{n_\lambda}, \mathbf{Z}_{n_\lambda}, \mathbf{N}_{1,n_\lambda} \in \mathbb{R}^{H\times W}$, $\odot$ denotes element-wise multiplication, and $\mathbf{N}_{1,n_\lambda}$ is the noise.

(2) Spectral dispersion ($\mathcal{D}$). The disperser $\mathcal{D}$ shifts each band along the H-axis by a band-dependent offset. Let $\mathbf{Z} \in \mathbb{R}^{H\times W\times N_\lambda}$ be the masked HSI and $\mathbf{Z}' \in \mathbb{R}^{(H+d(N_\lambda-1))\times W\times N_\lambda}$ the dispersed output, where d is the per-band shift. For each pixel (h, w, n_λ):

$$\mathbf{Z}'(h + d(n_\lambda - 1), w, n_\lambda) = \mathbf{Z}(h, w, n_\lambda) + n_2, \tag{3}$$

where n_2 denotes the noise at pixel (h, w) in spectral band n_λ. Thus, the dispersion operation on each spectral band can be defined as a circular shift matrix of dimension $\mathbb{R}^{H\times H}$ along the H-axis.

(3) Spectral integration ($\mathcal{S}$). The spectral integration operator $\mathcal{S}$ sums all dispersed bands, together with sensor noise $\mathbf{N}_3$, into a single 2D measurement $\mathbf{Y} \in \mathbb{R}^{(H+d(N_\lambda-1))\times W}$:

$$\mathbf{Y} = \mathcal{S}(\mathbf{Z}') = \sum\nolimits_{n_\lambda=1}^{N_\lambda} \mathbf{Z}'_{n_\lambda} + \mathbf{N}_3. \tag{4}$$

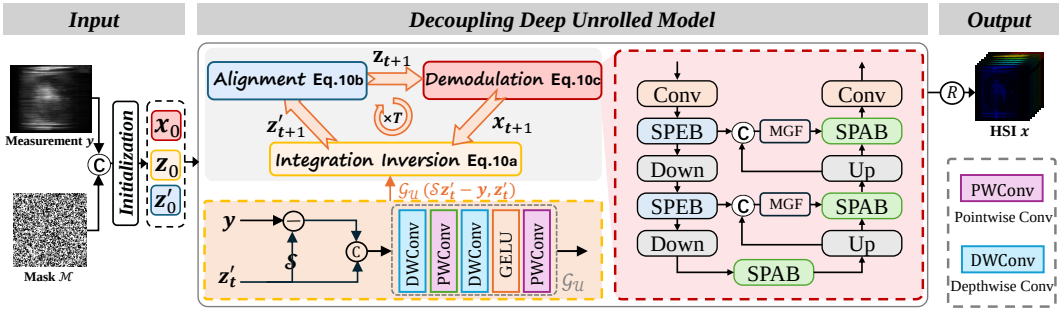


Figure 2: Overview of our Decoupled Deep Unrolling Model. Following the CASSI cascade, reconstruction is decomposed into three stages: integration inversion, spectral alignment, and mask demodulation. The orange box highlights the generalized physical inverse operator $\mathcal{G}_{\mathcal{U}}$ for integration inversion, while the red box denotes the proximal regularization network $\mathcal{D}_\theta$ for mask demodulation.

Coupled formulation and operator analysis. Most existing DUMs formulate CASSI reconstruction as a coupled inverse problem governed by the composite operator $\mathcal{A}$:

$$\mathbf{Y} = \mathcal{S}\mathcal{D}\mathcal{M}\,\mathbf{X} + \mathbf{N} \triangleq \mathcal{A}\,\mathbf{X} + \mathbf{N}, \tag{5}$$

where $\mathcal{M}$, $\mathcal{D}$, and $\mathcal{S}$ are the linear operators corresponding to eqs. (2) to (4), respectively.

To characterize the resulting inverse problem, we compare $\mathcal{A}$ with its constituent operators in table 1, with detailed derivations provided in the supplementary material. Compared with its constituent operators, $\mathcal{A}$ exhibits a lower information retention ratio, a higher-dimensional null space, and poorer conditioning. Together, these properties make reliable HSI recovery more challenging. Moreover, existing coupled formulations jointly address the inverse tasks associated with the individual imaging operations without dedicated regularization of the intermediate physical states. As a result, the distinct structural priors of these states are not explicitly exploited to guide reconstruction, potentially limiting the fidelity of the recovered HSI.

Motivated by these limitations, we formulate a decoupled reconstruction model that explicitly introduces intermediate state variables and decomposes CASSI reconstruction into operation-specific subproblems. This model exploits the structure of each physical operator and regularizes the corresponding intermediate states with dedicated learnable priors to improve HSI reconstruction fidelity.

3.2 Decoupled Reconstruction Model

Motivated by this analysis, we decouple the reconstruction by introducing physically meaningful intermediate variables along the CASSI imaging chain. Specifically, let $\mathbf{z}'$, $\mathbf{z}$, and $\mathbf{x}$ denote the pre-integration, pre-dispersion, and pre-modulation states, respectively. Under a Bayesian formulation, the joint posterior is factorized according to the CASSI imaging chain as

$$p(\mathbf{x}, \mathbf{z}, \mathbf{z}' \mid \mathbf{y}) \propto p(\mathbf{x})p(\mathbf{z})p(\mathbf{z}')p(\mathbf{z} \mid \mathbf{x})p(\mathbf{z}' \mid \mathbf{z})p(\mathbf{y} \mid \mathbf{z}'), \tag{6}$$

where the conditional factors follow the cascaded CASSI process $\mathbf{x} \to \mathbf{z} \to \mathbf{z}' \to \mathbf{y}$. We model the conditional likelihoods according to the operators in eqs. (2) to (4) with i.i.d. Gaussian noise:

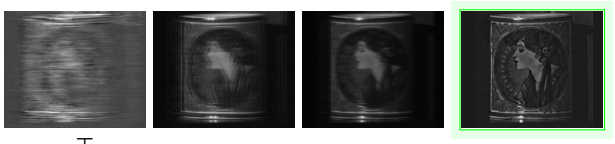
$$\mathbf{z} \sim \mathcal{N}(\mathcal{M}\mathbf{x}, \sigma_1^2\mathbf{I}), \quad \mathbf{z}' \sim \mathcal{N}(\mathcal{D}\mathbf{z}, \sigma_2^2\mathbf{I}), \quad \mathbf{y} \sim \mathcal{N}(\mathcal{S}\mathbf{z}', \sigma_3^2\mathbf{I}), \tag{7}$$

and further impose explicit Gibbs-form priors:

$$p(\mathbf{x}) \propto \exp(-\tau_1 R_1(\mathbf{x})),\ p(\mathbf{z}) \propto \exp(-\tau_2 R_2(\mathbf{z})),\ p(\mathbf{z}') \propto \exp(-\tau_3 R_3(\mathbf{z}')), \tag{8}$$

where $R_i(\cdot)$ denotes a learnable regularizer tailored to the physical characteristics of each latent state, and $\tau_i > 0$ controls its contribution. The MAP estimation of eq. (6) is equivalent to minimizing its negative log-posterior. Substituting the Gaussian likelihoods in eq. (7) and the Gibbs priors in eq. (8) yields the following decoupled objective:

$$\arg\min_{\mathbf{x},\mathbf{z},\mathbf{z}'} \tfrac{\lambda_1}{2}\|\mathbf{z} - \mathcal{M}\mathbf{x}\|_2^2 + \tfrac{\lambda_2}{2}\|\mathbf{z}' - \mathcal{D}\mathbf{z}\|_2^2 + \tfrac{\lambda_3}{2}\|\mathbf{y} - \mathcal{S}\mathbf{z}'\|_2^2 + \tau_1 R_1(\mathbf{x}) + \tau_2 R_2(\mathbf{z}) + \tau_3 R_3(\mathbf{z}'), \tag{9}$$



$\mathcal{S}^\top$ $\mathcal{G}_\mathcal{S}$ $\mathcal{G}_\mathcal{U}$ GT Gradient

Figure 3: Gradient maps of different operators vs. GT.

Table 2: Reconstruction quality comparison of different operators.

Operator	$\mathcal{S}^\top$	$\mathcal{G}_\mathcal{S}$	$\mathcal{G}_\mathcal{U}$
PSNR	40.96	41.10	41.43
Params	0.65	0.68	0.68

where $\lambda_i \propto 1/\sigma_i^2$. We optimize eq. (9) via alternating minimization, where each variable is updated in turn by solving its own subproblem with the others held fixed.

At iteration t, we sequentially update $\mathbf{z}'$, $\mathbf{z}$, and $\mathbf{x}$, yielding:

$$\begin{cases} \mathbf{z}'^{(t+1)} = \arg\min_{\mathbf{z}'} \mathcal{J}_1(z') = \frac{\lambda_3}{2}\|\mathbf{y} - \mathcal{S}\mathbf{z}'\|_2^2 + \frac{\lambda_2}{2}\left\|\mathbf{z}' - \mathcal{D}\mathbf{z}^{(t)}\right\|_2^2 + \tau_3 R_3(\mathbf{z}'), & \text{(10a)} \\ \mathbf{z}^{(t+1)} = \arg\min_{\mathbf{z}} \mathcal{J}_2(z) = \frac{\lambda_1}{2}\left\|\mathbf{z} - \mathcal{M}\mathbf{x}^{(t)}\right\|_2^2 + \frac{\lambda_2}{2}\left\|\mathbf{z}'^{(t+1)} - \mathcal{D}\mathbf{z}\right\|_2^2 + \tau_2 R_2(\mathbf{z}), & \text{(10b)} \\ \mathbf{x}^{(t+1)} = \arg\min_{\mathbf{x}} \mathcal{J}_3(x) = \frac{\lambda_1}{2}\left\|\mathbf{z}^{(t+1)} - \mathcal{M}\mathbf{x}\right\|_2^2 + \tau_1 R_1(\mathbf{x}). & \text{(10c)} \end{cases}$$

Compared with the coupled formulation in eq. (1), eq. (10) decomposes HSI recovery into three physically interpretable subproblems: *integration inversion*, *spectral alignment*, and *mask demodulation*, corresponding to the inverse of $\mathcal{S}$, $\mathcal{D}$, and $\mathcal{M}$, respectively. As characterized in table 1, each subproblem involves a better-conditioned operator than the coupled operator $\mathcal{A}$, making the corresponding subproblem substantially more tractable. More importantly, each latent state is explicitly constrained by a dedicated regularizer during reconstruction. This improves the accuracy of intermediate estimates and ultimately enhances HSI reconstruction fidelity.

3.3 Problem Solving

Integration Inversion for $\mathbf{z}'$. The $\mathbf{z}'$-subproblem in eq. (10a) aims to reconstruct the pre-integration cube $\mathbf{z}'$ from the single-channel measurement $\mathbf{y}$. As is standard in deep unrolling methods Ayad et al. (2024), we assume the regularization term R in eq. (10) to be differentiable. A simple gradient descent scheme can be applied to solve the optimization problem:

$$\mathbf{z}'^{(t+1)} = \mathbf{z}'^{(t)} - \lambda_3 \mathcal{S}^\top\big(\mathcal{S}\mathbf{z}'^{(t)} - \mathbf{y}\big) - \tau_3 \nabla R_3\big(\mathbf{z}'^{(t)}\big) - \lambda_2\big(\mathbf{z}'^{(t)} - \mathcal{D}\mathbf{z}^{(t)}\big). \tag{11}$$

where $\mathcal{S}^\top$ is the adjoint of the integration operator, which replicates the measurement residual across spectral channels. Although $\mathcal{S}^\top$ gives the gradient of the quadratic data-fidelity term, it does not explicitly distinguish the contributions of different spatially shifted spectral bands. As a result, the back-projected residual provides limited band-discriminative information for the highly underdetermined integration-inversion problem, which may reduce the effectiveness of iterative reconstruction.

We therefore introduce a learnable generalized physical inverse operator $\mathcal{G}_\mathcal{S}$, implemented as a learned back-projection operator, to augment the fixed adjoint $\mathcal{S}^\top$. Rather than simply replicating the mixed residual across spectral channels, $\mathcal{G}_\mathcal{S}$ learns band-dependent corrections to better separate the contributions of spatially shifted spectral responses. The update rule is formulated as:

$$\mathbf{z}'^{(t+1)} = \mathbf{z}'^{(t)} - \lambda_3 \mathcal{G}_\mathcal{S}\big(\mathcal{S}\mathbf{z}'^{(t)} - \mathbf{y}\big) - \tau_3 \nabla R_3\big(\mathbf{z}'^{(t)}\big) - \lambda_2\big(\mathbf{z}'^{(t)} - \mathcal{D}\mathbf{z}^{(t)}\big), \tag{12}$$

Compared with the fixed adjoint back-projection $\mathcal{S}^\top$, the learnable $\mathcal{G}_\mathcal{S}$ produces more informative band-aware update directions, leading to more accurate and stable reconstruction (fig. 3 and table 2).

To condition this correction on the current reconstruction, we further combine learned residual back-projection and state-specific regularization in a unified operator $\mathcal{G}_\mathcal{U}$. Inspired by the synergy of joint learning Xin et al. (2024); Guo et al. (2025), $\mathcal{G}_\mathcal{U}$ takes the physically consistent residual $(\mathcal{S}\mathbf{z}'^{(t)} - \mathbf{y})$ and $\mathbf{z}'^{(t)}$ as inputs and predicts their joint contribution to the update:

$$\mathbf{z}'^{(t+1)} = \mathbf{z}'^{(t)} - \lambda_3\, \mathcal{G}_\mathcal{U}\big(\mathcal{S}\mathbf{z}'^{(t)} - \mathbf{y},\, \mathbf{z}'^{(t)}\big) - \lambda_2(\mathbf{z}'^{(t)} - \mathcal{D}\mathbf{z}^{(t)}), \tag{13}$$

This joint formulation enables the residual information and the current reconstruction state to jointly determine the update direction. As shown in fig. 3 and table 2, $\mathcal{G}_\mathcal{U}$ provides more effective updates and further improves reconstruction performance.

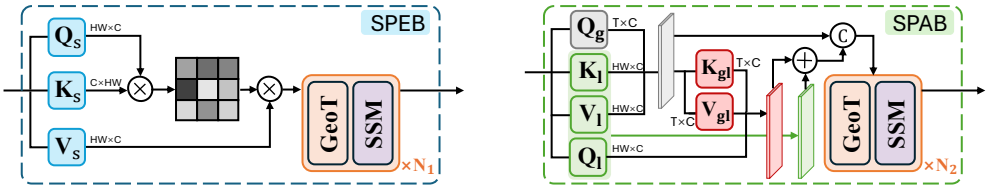


Figure 4: Overview of the proposed Spectral Attention Block (SPEB) and Spatial Attention Block (SPAB). GeoT denotes the geometric transformation operation.

Design of $\mathcal{G}_{\mathcal{U}}$. Due to spectral dispersion, the bands of $\mathbf{z}'$ are shifted by wavelength-dependent offsets, while spectral integration superimposes their contributions at each detector coordinate to form the measurement $\mathbf{y}$. Reconstructing $\mathbf{z}'$ therefore requires separating the mixed spectral contributions while preserving the dispersed geometry for subsequent spectral alignment. To this end, $\mathcal{G}_{\mathcal{U}}$ (see Fig. 2) combines pointwise convolutions to model cross-channel dependencies with depthwise convolutions to aggregate local spatial context. The resulting module predicts band-dependent corrections that refine the current estimate of $\mathbf{z}'$ on the dispersed grid.

Spectral Alignment for z. The $\mathbf{z}$-subproblem reconstructs the pre-dispersion cube $\mathbf{z}$ from its dispersed state $\mathbf{z}'$. In the CASSI system, the dispersion operator $\mathcal{D}$ is a known and invertible shift operator satisfying $\mathcal{D}^{-1} = \mathcal{D}^{\top}$. Thus, the eq. (10b) admits a closed-form update by setting $\nabla_{\mathbf{z}}\mathcal{J}_2(\mathbf{z}) = 0$:

$$\lambda_1(\mathbf{z} - \mathcal{M}\mathbf{x}^{(t)}) + \lambda_2\mathcal{D}^{\top}(\mathcal{D}\mathbf{z} - \mathbf{z}'^{(t+1)}) + \tau_2\,\nabla R_2(\mathbf{z}^{(t)}) = 0. \tag{14}$$

Using $\mathcal{D}^{\top}\mathcal{D} = \mathbf{I}$, we obtain

$$\mathbf{z}^{(t+1)} = (\lambda_1\,\mathcal{M}\mathbf{x}^{(t)} + \lambda_2\,\mathcal{D}^{\top}\mathbf{z}'^{(t+1)} - \tau_2\,\nabla R_2(\mathbf{z}^{(t)}))/(\lambda_1 + \lambda_2). \tag{15}$$

Although the spectral dispersion is readily invertible, it introduces wavelength-dependent spatial shifts across spectral bands in $\mathbf{z}'$. Therefore, the prior R_2 is designed to correct the local spatial misalignment induced by dispersion. We implement $\nabla R_2(\cdot)$ with a shallow convolutional network, whose local receptive fields are well suited to modeling spatial correspondence across shifted bands.

Mask Demodulation for x. The $\mathbf{x}$-subproblem in eq. (10c) recovers the latent HSI $\mathbf{x}$ from the modulated cube $\mathbf{z}^{(t+1)}$ while enforcing the prior $R_1(\mathbf{x})$. To decouple data fidelity from prior modeling, we introduce an auxiliary variable $\mathbf{v}$ and rewrite the problem as:

$$\begin{cases} \mathbf{v}^{(t)} = \arg\min_{\mathbf{v}} \frac{\eta_1}{2}\left\|\mathbf{v} - \mathbf{x}^{(t)}\right\|_2^2 + \frac{\lambda_1}{2}\left\|\mathbf{z}^{(t+1)} - \mathcal{M}\mathbf{v}\right\|_2^2, & \text{(16a)} \\ \mathbf{x}^{(t+1)} = \arg\min_{\mathbf{x}} \frac{\eta_1}{2}\left\|\mathbf{v}^{(t)} - \mathbf{x}\right\|_2^2 + \tau_1 R_1(\mathbf{x}). & \text{(16b)} \end{cases}$$

eq. (16a) is a standard quadratic problem with a closed-form solution:

$$\mathbf{v}^{(t)} = (\lambda_1\mathcal{M}^{\top}\mathcal{M} + \eta_1\mathbf{I})^{-1}\big(\lambda_1\mathcal{M}^{\top}\mathbf{z}^{(t+1)} + \eta_1\mathbf{x}^{(t)}\big), \tag{17}$$

where $\eta_1 > 0$ is a stabilizing parameter and $\mathcal{M}^{\top} = \mathcal{M}$. The second subproblem corresponds to a proximal step for enforcing the prior $R_1(\mathbf{x})$. We implement this proximal map with a dedicated regularization network $\mathcal{D}_\theta$:

$$\mathbf{x}^{(t+1)} = \mathrm{ProxNet}_x\big(\mathbf{v}^{(t+1)}\big) \;\equiv\; \mathcal{D}_\theta\big(\mathbf{v}^{(t+1)}\big), \tag{18}$$

Design of $\mathcal{D}_\theta$. The regularization network $\mathcal{D}_\theta$ learns spatial–spectral priors to guide HSI recovery during mask demodulation. We adopt an asymmetric spatial–spectral U-Net, with spectral attention blocks in the encoder to model cross-channel dependencies, and spatial attention blocks in the bottleneck and decoder to aggregate spatial context and refine structural details. Both block types employ lightweight Mamba modules, each preceded by a geometric transformation (GeoT) that reorients the input features according to a predefined cyclic schedule, enabling successive modules to capture dependencies along different scanning directions. To incorporate the coded aperture into feature propagation, we introduce mask-guided fusion (MGF) in the skip connections. At each scale, the spatially aligned mask is resized and encoded by a lightweight convolutional branch. The resulting mask embedding is concatenated with the encoder and decoder features, followed by a convolutional predictor and sigmoid activation to generate a gating map. The gate multiplicatively modulates the encoder features, yielding mask- and context-aware skip features for subsequent decoder fusion. Compared with direct skip forwarding, MGF adaptively filters encoder features according to the physical mask and current reconstruction context (see Appendix for details).

Table 3: PSNR (dB, top) and SSIM (bottom) on 10 synthetic spectral scenes.

Algorithms	Params(M)	FLOPs(G)	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg
TSA-Net Meng et al. (2020)	44.25	110.06	32.03 0.892	31.00 0.858	32.25 0.915	39.19 0.953	29.39 0.884	31.44 0.908	30.32 0.878	29.35 0.888	30.01 0.890	29.59 0.874	31.46 0.894
DAUHST Cai et al. (2022d)	6.15	79.50	37.25 0.958	39.02 0.967	41.05 0.971	46.15 0.983	35.80 0.969	37.08 0.970	37.57 0.963	35.10 0.966	40.02 0.970	34.59 0.956	38.36 0.967
BIRNAT Cheng et al. (2023)	4.40	212.55	36.78 0.951	37.89 0.957	40.61 0.971	46.93 0.985	35.42 0.963	35.30 0.959	36.58 0.954	33.95 0.955	39.46 0.969	32.80 0.937	37.57 0.960
GAP-Net Meng et al. (2023)	4.27	78.58	33.63 0.913	33.19 0.902	33.96 0.931	39.14 0.971	31.44 0.921	32.29 0.927	31.79 0.903	30.25 0.907	33.06 0.916	30.14 0.898	32.89 0.919
RDLUF-MixS^2 Dong et al. (2023)	1.89	115.34	37.94 0.966	40.95 0.977	43.25 0.979	47.83 0.990	37.11 0.976	37.47 0.975	38.58 0.969	35.50 0.970	41.83 0.978	35.23 0.962	39.57 0.974
DiffSCI Pan et al. (2024)	93.52	193.76	34.96 0.907	34.60 0.905	39.83 0.949	42.65 0.951	35.21 0.946	33.12 0.917	36.29 0.944	30.42 0.887	37.27 0.931	28.49 0.821	35.28 0.916
DADF-Plus-3 Xu et al. (2024)	58.13	230.41	37.46 0.965	39.86 0.976	41.03 0.974	45.98 0.989	35.53 0.972	37.02 0.975	36.76 0.958	34.78 0.971	40.07 0.976	34.39 0.962	38.29 0.972
DERNN Dong et al. (2024)	0.62	81.99	38.26 0.965	40.97 0.979	43.22 0.979	48.10 0.991	38.08 0.980	37.41 0.975	38.83 0.971	36.41 0.973	42.87 0.981	35.15 0.962	39.93 0.976
DPU Zhang et al. (2024a)	2.85	49.26	38.91 0.968	41.99 0.981	44.10 0.980	48.33 0.990	38.07 0.978	38.58 0.978	39.13 0.971	36.90 0.975	42.88 0.981	36.36 0.970	40.52 0.977
SSR-L Zhang et al. (2024b)	5.18	78.93	39.07 0.970	42.04 0.981	44.49 0.980	48.80 0.990	38.64 0.980	38.50 0.978	39.16 0.971	36.96 0.976	43.12 0.982	36.08 0.968	40.69 0.978
LADE Wu et al. (2025)	2.78	88.68	38.08 0.969	41.84 0.982	43.77 0.983	47.99 0.993	37.97 0.980	38.30 0.980	38.82 0.973	36.15 0.979	42.53 0.984	35.48 0.970	40.09 0.979
MiJUN Qin et al. (2025)	0.56	73.67	39.26 0.973	41.78 0.983	44.31 0.983	48.53 0.994	39.30 0.985	38.22 0.979	41.00 **0.983**	36.72 0.978	43.84 0.985	35.56 0.967	40.86 0.982
LRDUN Huang et al. (2026)	2.04	30.58	**39.45** 0.973	40.97 0.979	43.63 0.982	48.76 0.993	39.36 0.984	**38.99** 0.980	40.24 0.979	**38.28** 0.981	43.39 0.985	**36.52** 0.973	40.96 0.982
Ours	0.68	72.17	39.36 **0.977**	**42.23** **0.986**	**45.21** **0.986**	**49.53** **0.995**	**40.13** **0.988**	38.65 **0.983**	**41.37** **0.983**	37.12 **0.983**	**44.43** **0.988**	36.25 **0.975**	**41.43** **0.984**

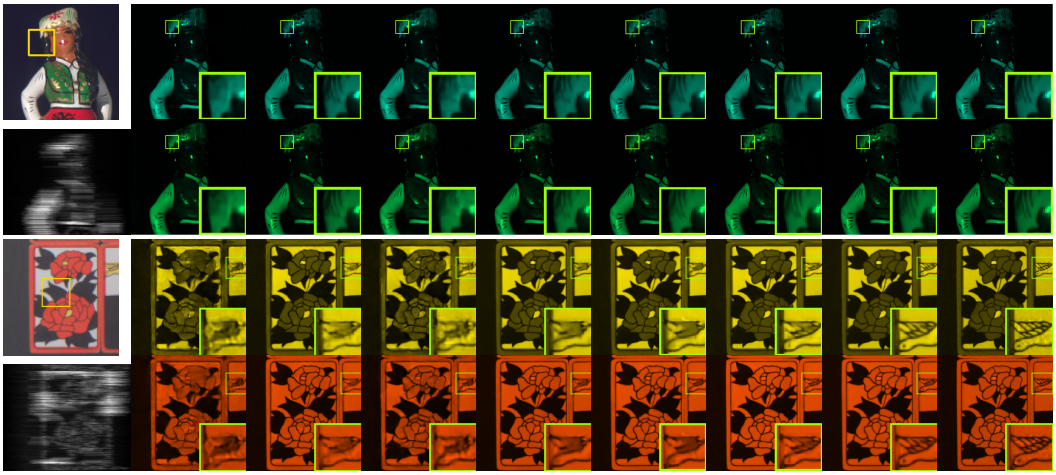


Figure 5: The simulated reconstruction results for Scene 6 (top) & Scene 7 (bottom) with 2 out of 28 spectral channels. The left displays the RGB image and measurement **Y**.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate DDUM on both synthetic and real CASSI measurements. Following Qin et al. (2025), we use CAVE Park et al. (2007) and KAIST Choi et al. (2017) for synthetic experiments. CAVE contains 32 hyperspectral images at 512×512 resolution, while KAIST contains 30 images at 2704×3376 resolution. We train on CAVE and use 10 representative KAIST scenes for testing. For real-world evaluation, we use the five real CASSI datasets from Meng et al. (2020).

Implementation Details. Following the training protocol of Cai et al. (2022d), we optimize the model with Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$) for 300 epochs under a cosine-annealing learning-rate schedule. The parameters λ_i and τ_i in eq. (9) are learnable and jointly optimized during training.

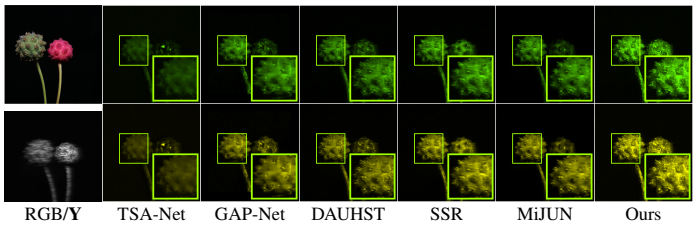


Figure 6: Visual comparison on the real dataset.

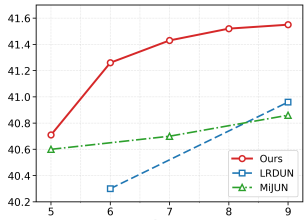


Figure 7: PSNR vs. Iteration.

Table 4: PSNR on the ICVL dataset.

Method	hill 0325	Flower 0325	eve 0331	pepper 0503	prk 0328
DERNN	37.72	46.38	44.76	38.14	38.18
LADE	37.20	48.12	43.32	41.13	39.69
MiJUN	37.47	47.56	45.96	40.12	40.04
Ours	**38.76**	**48.52**	**46.33**	**41.20**	**41.17**

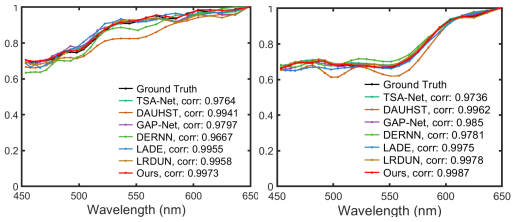


Figure 8: Spatial–spectral correlation profiles.

4.2 Comparison with State-of-the-Art

Synthetic Data. table 3 reports quantitative results on ten synthetic scenes. DDUM achieves the highest PSNR of 41.43 dB, demonstrating the effectiveness of explicitly decoupling the cascaded CASSI reconstruction process. Compared with MiJUN, DDUM achieves higher reconstruction quality with fewer FLOPs and a comparable parameter budget, indicating that the performance gain arises from the proposed reconstruction formulation rather than increased model complexity.

The visual comparisons in fig. 5 further confirm the reconstruction fidelity of DDUM. As shown in the hat textures in the top example and feather details in the bottom one, DDUM preserves fine structures better, while existing methods over-smooth these areas or introduce artifacts.

Real Data. To evaluate robustness to real-world acquisition conditions, we evaluate the methods on real CASSI measurements with noise. Since ground-truth hyperspectral cubes are unavailable, we provide qualitative comparisons in fig. 6. All models are trained on synthetic data. DDUM produces clearer floral structures with fewer noise residues and artifacts, while existing methods exhibit more pronounced blurring and noise amplification under real measurement conditions.

Spatial–Spectral Fidelity. We further evaluate spatial–spectral fidelity by comparing the profiles of two selected regions (yellow boxes) in fig. 5, with the corresponding spectral curves shown in fig. 8. DDUM achieves the highest correlation with the ground truth and most faithfully preserves the spectral-curve shapes. This improved fidelity indicates that the proposed decoupled reconstruction better preserves both spatial structures and spectral responses throughout the reconstruction process.

Iteration Analysis. Fig. 7 compares DDUM with existing models across iteration counts. DDUM reaches 41.43 dB with only 8 iterations, already outperforming coupled models with more iterations. Moreover, at the same 9 iterations, DDUM achieves a 0.59 dB gain over MiJUN, demonstrating that the decoupled formulation enables more effective reconstruction at each iteration.

Cross-Scene Generalization. To evaluate cross-scene generalization, we test all methods on unseen natural scenes from the ICVL dataset Arad & Ben-Shahar (2016). As reported in table 4, DDUM maintains a clear performance advantage on these out-of-distribution scenes, demonstrating that the proposed decoupled formulation generalizes beyond the training distribution.

4.3 Ablation Study

Effectiveness of the decoupling paradigm. To evaluate the proposed decoupled paradigm, we keep the architectures of several representative CASSI models unchanged and replace only their coupled reconstruction with our decoupled formulation. As shown in table 5, all models achieve consistent performance gains with only minor increases in parameters and computation, indicating that

Table 5: Ablation study of the decoupling paradigm and intermediate-state priors. Left: coupled vs. decoupled formulations; right: different prior configurations. † denotes a stronger R_1 network.

Decoupling Study				**Prior Study**				
Method	Params	FLOPs	PSNR	$R_1(\mathbf{x})$	$R_2(\mathbf{z})$	$R_3(\mathbf{z}')$	Params	PSNR
DAUHST	6.15 / 6.18	79.50 / 81.72	38.36 / **39.11**	✓			0.57	40.84
DERNN	0.62 / 0.64	81.99 / 83.21	39.93 / **40.51**	✓†			0.72	40.85
MiJUN	0.56 / 0.59	73.67 / 75.90	40.86 / **41.11**	✓	✓		0.68	41.03
Ours	0.65 / 0.68	70.20 / 72.17	40.84 / **41.43**	✓	✓	✓	0.68	**41.43**

Table 6: Noise robustness.

Method	PSNR	Drop
DERNN	39.23	0.70
LADE	39.20	0.89
MiJUN	40.44	0.42
Ours	**41.21**	**0.22**

Table 7: Ablation of $\mathcal{D}_\theta$.

w/o	PSNR	Drop
SPEB	41.31	0.12
SPAB	41.19	0.24
MGF	41.28	0.15
Ours	**41.43**	–

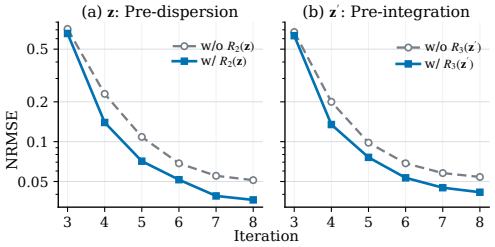


Figure 9: Reconstruction errors of $\mathbf{z}$ and $\mathbf{z}'$.

the improvements mainly stem from the reconstruction reformulation rather than increased model capacity. These results further demonstrate the general applicability of our decoupling paradigm to existing unrolling networks.

Role of intermediate-state priors. To evaluate the role of intermediate-state priors, we fix $R_1(\mathbf{x})$ as the prior on the final HSI and progressively introduce $R_2(\mathbf{z})$ and $R_3(\mathbf{z}')$. As shown in fig. 9, introducing these priors substantially reduces the reconstruction errors of $\mathbf{z}$ and $\mathbf{z}'$, indicating more accurate intermediate-state estimation. Since DDUM reconstructs the intermediate states sequentially following the CASSI imaging process, these improvements provide more reliable inputs for subsequent stages and ultimately benefit the final HSI reconstruction. As shown in table 5, using only $R_1(\mathbf{x})$ yields 40.84 dB, while adding $R_2(\mathbf{z})$ and $R_3(\mathbf{z}')$ further improves PSNR by 0.19 dB and 0.40 dB, respectively, demonstrating the effectiveness of explicitly regularizing intermediate states.

Noise robustness. We further evaluate the robustness of different methods by adding Gaussian noise with $\sigma = 0.05$ to the normalized measurements, as shown in table 6. DDUM achieves the highest PSNR of 41.21 dB under noise and the smallest drop of only 0.22 dB, compared with 0.70, 0.89, and 0.42 dB for DERNN, LADE, and MiJUN, respectively. These results demonstrate that DDUM maintains more stable reconstruction performance in the presence of measurement noise.

Ablation of $\mathcal{D}_\theta$. We ablate the main components of the $\mathcal{D}_\theta$ in table 7. Removing SPEB, SPAB, and MGF decreases the PSNR by 0.12, 0.24, and 0.15 dB, respectively, confirming that all three components contribute to the final reconstruction. In particular, the larger degradation without SPAB highlights the importance of spatial refinement, while the performance drop without MGF verifies the benefit of incorporating mask-guided feature fusion into the skip connections.

5 Conclusion

We presented a Decoupled Deep Unrolling Model for CASSI reconstruction, which reformulates the conventional coupled inverse problem into three tractable and physically interpretable inverse subproblems. By introducing explicit intermediate states and stage-specific regularization mechanisms, DDUM enables targeted reconstruction at each inverse stage and improves the fidelity of the recovered hyperspectral images. Extensive experiments on multiple benchmark datasets demonstrate that DDUM consistently outperforms existing deep unrolling methods in both reconstruction quality and computational efficiency. Moreover, existing state-of-the-art unrolling networks can be seamlessly incorporated into our framework to further improve performance. Despite its advantages, DDUM has been evaluated primarily under standard CASSI configurations. Exploring robustness to real-world hardware imperfections is a promising direction for future work.

REFERENCES

Boaz Arad and Ohad Ben-Shahar. Sparse recovery of hyperspectral signal from natural RGB images. In *European Conference on Computer Vision*, pp. 19–34. Springer, 2016.

Ishak Ayad, Nicolas Larue, and Maï K Nguyen. QN-Mixer: A quasi-newton MLP-Mixer model for sparse-view CT reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 25317–25326, 2024. doi: 10.1109/CVPR52733.2024.02392.

Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Coarse-to-fine sparse transformer for hyperspectral image reconstruction. In *European Conference on Computer Vision*, pp. 686–704. Springer, 2022a.

Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17502–17511, 2022b.

Yuanhao Cai, Jing Lin, Zudi Lin, Haoqian Wang, Yulun Zhang, Hanspeter Pfister, Radu Timofte, and Luc Van Gool. MST++: Multi-stage spectral-wise transformer for efficient spectral reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 745–755, 2022c.

Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc Van Gool. Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging. *Advances in Neural Information Processing Systems*, 35:37749–37761, 2022d.

Ziheng Cheng, Bo Chen, Ruiying Lu, Zhengjue Wang, Hao Zhang, Ziyi Meng, and Xin Yuan. Recurrent neural networks for snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2264–2281, 2023. doi: 10.1109/TPAMI.2022.3161934.

Inchang Choi, Daniel S. Jeon, Giljoo Nam, Diego Gutierrez, and Min H. Kim. High-quality hyperspectral reconstruction using a spectral prior. *ACM Transactions on Graphics*, 36(6):1–13, 2017.

Aram Danielyan, Vladimir Katkovnik, and Karen Egiazarian. BM3D frames and variational image deblurring. *IEEE Transactions on image processing*, 21(4):1715–1728, 2011.

Yubo Dong, Dahua Gao, Tian Qiu, Yuyan Li, Minxi Yang, and Guangming Shi. Residual degradation learning unfolding framework with mixing priors across spectral and spatial for compressive spectral imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22262–22271, 2023.

Yubo Dong, Dahua Gao, Yuyan Li, Guangming Shi, and Danhua Liu. Degradation estimation recurrent neural network with local and non-local priors for compressive spectral imaging. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024. doi: 10.1109/TGRS.2024.3410272.

Bo Du, Yuxiang Zhang, Liangpei Zhang, and Dacheng Tao. Beyond the sparsity-based target detector: A hybrid sparsity and statistics-based detector for hyperspectral images. *IEEE Transactions on Image Processing*, 25(11):5345–5357, 2016.

Shuang Guo, Friedhelm Hamann, and Guillermo Gallego. Unsupervised joint learning of optical flow and intensity with event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7980–7989, 2025. doi: 10.1109/ICCV51701.2025.00748.

He Huang, Yujun Guo, and Wei He. LRDUN: A low-rank deep unfolding network for efficient spectral compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10556–10566, June 2026.

Tao Huang, Weisheng Dong, Xin Yuan, Jinjian Wu, and Guangming Shi. Deep gaussian scale mixture prior for spectral compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16216–16225, 2021.

Yixing Ji, Shengjiang Kong, Weiwei Wang, Xixi Jia, and Xiangchu Feng. Iterative decoupling deconvolution network for image restoration. *Journal of Visual Communication and Image Representation*, 104:104288, 2024.

Emmanuel Maggiori, Guillaume Charpiat, Yuliya Tarabalka, and Pierre Alliez. Recurrent neural networks to correct satellite image classification maps. *IEEE Transactions on Geoscience and Remote Sensing*, 55(9):4962–4971, 2017.

Ziyi Meng, Jiawei Ma, and Xin Yuan. End-to-end low cost compressive spectral imaging with spatial-spectral self-attention. In *European conference on computer vision*, pp. 187–204. Springer, 2020.

Ziyi Meng, Xin Yuan, and Shirin Jalali. Deep unfolding for snapshot compressive imaging. *International Journal of Computer Vision*, 131(11):2933–2958, 2023.

Xin Miao, Xin Yuan, Yunchen Pu, and Vassilis Athitsos. λ-Net: Reconstruct hyperspectral images from a snapshot measurement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4059–4069, 2019. doi: 10.1109/ICCV.2019.00416.

Zhenghao Pan, Haijin Zeng, Jiezhang Cao, Kai Zhang, and Yongyong Chen. DiffSCI: Zero-shot snapshot compressive imaging via iterative spectral diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23901–23911, June 2024. doi: 10.1109/CVPR52733.2024.02390.

Jong-Il Park, Moon-Hyun Lee, Michael D Grossberg, and Shree K Nayar. Multispectral imaging using multiplexed illumination. In *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8. IEEE, 2007.

Mengjie Qin, Yuchao Feng, Zongliang Wu, Yulun Zhang, and Xin Yuan. Detail matters: Mamba-inspired joint unfolding network for snapshot spectral compressive imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6594–6602, 2025.

Lizhi Wang, Chen Sun, Ying Fu, Min H Kim, and Hua Huang. Hyperspectral image reconstruction using a deep spatial-spectral prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8032–8041, 2019.

Lizhi Wang, Chen Sun, Maoqing Zhang, Ying Fu, and Hua Huang. DNU: Deep non-local unrolling for computational spectral imaging. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1658–1668, 2020a. doi: 10.1109/CVPR42600.2020.00173.

Qian Wang, Li Sun, Yan Wang, Mei Zhou, Menghan Hu, Jiangang Chen, Ying Wen, and Qingli Li. Identification of melanoma from hyperspectral pathology image using 3D convolutional networks. *IEEE Transactions on Medical Imaging*, 40(1):218–227, 2020b.

Xiaodong Wang, Zijun He, Ping Wang, Lishun Wang, Yanan Hu, and Xin Yuan. Spectral compressive imaging via chromaticity-intensity decomposition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Zongliang Wu, Ruiying Lu, Ying Fu, and Xin Yuan. Latent diffusion prior enhanced deep unfolding for snapshot spectral compressive imaging. In *European Conference on Computer Vision*, pp. 164–181. Springer, 2025.

Bingyu Xin, Meng Ye, Leon Axel, and Dimitris N Metaxas. Rethinking deep unrolled model for accelerated MRI reconstruction. In *European Conference on Computer Vision*, pp. 164–181. Springer, 2024.

Ping Xu, Lei Liu, Haifeng Zheng, Xin Yuan, Chen Xu, and Lingyun Xue. Degradation-aware dynamic Fourier-based network for spectral compressive imaging. *IEEE Transactions on Multimedia*, 26:2838–2850, 2024. doi: 10.1109/TMM.2023.3304450.

Zhiyang Yao, Shuyang Liu, Xiaoyun Yuan, and Lu Fang. SPECAT: Spatial-spectral cumulative-attention transformer for high-resolution hyperspectral image reconstruction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25368–25377, 2024. doi: 10.1109/CVPR52733.2024.02397.

Xin Yuan. Generalized alternating projection based total variation minimization for compressive sensing. In *2016 IEEE International conference on image processing*, pp. 2539–2543, 2016.

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

Jiancheng Zhang, Haijin Zeng, Jiezhang Cao, Yongyong Chen, Dengxiu Yu, and Yin-Ping Zhao. Dual prior unfolding for snapshot compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25742–25752, June 2024a. doi: 10.1109/CVPR52733.2024.02432.

Jiancheng Zhang, Haijin Zeng, Yongyong Chen, Dengxiu Yu, and Yin-Ping Zhao. Improving spectral snapshot reconstruction with spectral-spatial rectification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25817–25826, June 2024b.

STATEMENT ON THE USE OF LARGE LANGUAGE MODELS

In this work, we used large language models (LLMs) as an auxiliary tool. The primary role of LLMs was to assist with language polishing, phrasing improvements, grammar checking, and enhancing the clarity and fluency of certain sentences, ensuring that our research ideas and technical details are expressed more accurately and professionally. Importantly, LLMs were not involved in the core aspects of this work, including research conception, result generation, or data analysis.

A ALGEBRAIC DERIVATION OF TAB. 1

This section presents the detailed algebraic derivations underlying the entries reported in Tab. 1 of the main text, including information retention ratio (IRR), null-space dimension (NullDim), observable-subspace condition number κ_{obs}, and the computational cost associated with solving the regularized normal equations for each CASSI operator. The derivations follow the modeling assumptions of the main text; for completeness, we restate the core notation below.

A.1 NOTATION

Let $n = H \times W$ denote the number of spatial samples per spectral band, and let N_λ denote the number of spectral channels. The unknown hyperspectral image (HSI) can be stacked bandwise as

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_{N_\lambda} \end{bmatrix} \in \mathbb{R}^{N_\lambda n}, \qquad x_k \in \mathbb{R}^n, \tag{19}$$

and the 2D CASSI measurement is given by $y \in \mathbb{R}^n$.

The CASSI forward operator consists of mask modulation ($\mathcal{M}$), spectral dispersion ($\mathcal{D}$), and spectral integration ($\mathcal{S}$), which take the following matrix forms:

$$\mathcal{M} = I_{N_\lambda} \otimes \operatorname{diag}(m) \in \mathbb{R}^{N_\lambda n \times N_\lambda n}, \tag{20}$$

$$\mathcal{D} = \operatorname{blockdiag}(P_1, \ldots, P_{N_\lambda}) \in \mathbb{R}^{N_\lambda n \times N_\lambda n}, \tag{21}$$

$$\mathcal{S} = \mathbf{1}_{1 \times N_\lambda} \otimes I_n = [\, I_n, \ldots, I_n \,] \in \mathbb{R}^{n \times N_\lambda n}, \tag{22}$$

where $m \in \{0,1\}^n$ is the coded aperture mask, and P_k is the permutation matrix corresponding to the dispersive shift of the k-th wavelength. The full CASSI forward operator is therefore written as:

$$\mathcal{A} = \mathcal{S}\mathcal{D}\mathcal{M} \in \mathbb{R}^{n \times N_\lambda n}. \tag{23}$$

A.2 INFORMATION RETENTION RATIO (IRR)

For a linear operator $\mathcal{T} : \mathbb{R}^{n_{\text{in}}} \to \mathbb{R}^{n_{\text{out}}}$, we define the information retention ratio (IRR) with respect to an input $x_{\text{in}} \in \operatorname{dom}(\mathcal{T})$ as the fraction of independent scalar degrees of freedom preserved after

applying $\mathcal{T}$:

$$\mathrm{IRR}(\mathcal{T}; x_{\mathrm{in}}) := \frac{n_{\mathrm{eff}}(\mathcal{T} x_{\mathrm{in}})}{n_{\mathrm{eff}}(x_{\mathrm{in}})}, \tag{24}$$

where $n_{\mathrm{eff}}(\cdot)$ denotes the number of independent scalar elements (pixels) in a vector. A smaller IRR indicates stronger compression, which reflects a higher difficulty in reconstruction. We now derive the IRR for each stage of the CASSI pipeline.

Modulation ($\mathcal{M}$). Let p denote the fraction of nonzero entries in $\mathrm{diag}(m)$. In each of the N_λ spectral channels with n spatial pixels, the mask preserves pn elements per channel. Therefore, the total number of retained elements is $pN_\lambda n$, and the corresponding IRR is:

$$\mathrm{IRR}(\mathcal{M}) = \frac{pN_\lambda n}{N_\lambda n} = p. \tag{25}$$

Dispersion ($\mathcal{D}$). The matrix $\mathcal{D} = \mathrm{blockdiag}(P_1, \ldots, P_{N_\lambda})$ contains one permutation matrix P_k in each spectral block. Since the permutation merely reorders entries without discarding information, so:

$$\mathrm{IRR}(\mathcal{D}) = 1. \tag{26}$$

Spectral integration ($\mathcal{S}$). At each spatial location, the integration operator $\mathcal{S} = [I_n, \ldots, I_n]$ sums all N_λ spectral channels into a single scalar. Consequently, only n output elements remain from the original $N_\lambda n$, giving

$$\mathrm{IRR}(\mathcal{S}) = \frac{n}{N_\lambda n} = \frac{1}{N_\lambda}. \tag{27}$$

Coupled operator ($\mathcal{A} = \mathcal{S}\mathcal{D}\mathcal{M}$). The coupled operator $\mathcal{A}$ first masks the HSI, preserving a fraction p of the $N_\lambda n$ input elements. Dispersion then permutes these surviving $pN_\lambda n$ elements without altering their count, and the subsequent spectral integration compresses each spatial group of N_λ spectral samples into a single scalar. Thus, the operator $\mathcal{A}$ retains roughly pn independent degrees of freedom:

$$\mathrm{IRR}(\mathcal{A}) \approx \mathrm{IRR}(\mathcal{S})\,\mathrm{IRR}(\mathcal{D})\,\mathrm{IRR}(\mathcal{M}) = \frac{pn}{N_\lambda n} = \frac{p}{N_\lambda}. \tag{28}$$

This formulation shows that the IRR of the coupled CASSI operator is approximately the product of the IRRs of its constituent operators, and is strictly smaller the IRR of each of the sub-operators of $\mathcal{A}$.

A.3 Null-space dimension (NullDim)

The null-space dimension of a linear operator is the dimension of its kernel. For the CASSI components, we obtain the following results.

Modulation ($\mathcal{M}$). The kernel of $\mathcal{M}$ comprises all vectors whose support lies entirely on the masked spatial locations, i.e., indices where $m_i = 0$. Each masked pixel contributes N_λ spectral degrees of freedom, and a fraction $(1 - p)$ of the n spatial positions are masked. Therefore, the null-space dimension is:

$$\dim \ker(\mathcal{M}) = (1 - p)N_\lambda n. \tag{29}$$

Dispersion ($\mathcal{D}$). Each block P_k in $\mathcal{D}$ is an invertible permutation matrix and thus induces a bijective reordering of the corresponding spectral band. Since no degrees of freedom are removed under such permutations, the dispersion operator has a trivial kernel:

$$\dim \ker(\mathcal{D}) = 0. \tag{30}$$

Spectral integration ($\mathcal{S}$). At each spatial location, $\mathcal{S}$ maps an N_λ-dimensional spectral vector to a single summed value. The nullity of this linear map is therefore $N_\lambda - 1$ per pixel, yielding:

$$\dim \ker(\mathcal{S}) = (N_\lambda - 1)n. \tag{31}$$

Coupled operator ($\mathcal{A} = \mathcal{SDM}$). Since $\operatorname{rank}(\mathcal{A}) \le \min(\operatorname{rank}(\mathcal{S}), \operatorname{rank}(\mathcal{D}), \operatorname{rank}(\mathcal{M})) \le \operatorname{rank}(\mathcal{S}) = n$, the kernel-space dimension of $\mathcal{A}$ satisfies:

$$\dim \ker(\mathcal{A}) = N_\lambda n - \operatorname{rank}(\mathcal{A}) \ge N_\lambda n - n = (N_\lambda - 1)n. \tag{32}$$

When $p < 1$, modulation fully suppresses a subset of spatial locations. Each such pixel contributes an additional N_λ-dimensional null subspace that is not captured by $\ker(\mathcal{S})$, strictly enlarging the kernel:

$$\dim \ker(\mathcal{A}) > (N_\lambda - 1)n. \tag{33}$$

This establishes the entry "NullDim $> (N_\lambda - 1)n$" in Tab. 1.

A.4 Observable Condition Number κ_{obs}

We next analyze the observable condition number κ_{obs} for each component of the CASSI forward model.

Modulation ($\mathcal{M}$). Restricted to the observable subspace (coordinates corresponding to unmasked pixels), $\mathcal{M}$ reduces to an identity matrix. Consequently, all singular values in this subspace are $\sigma_{\max}(\mathcal{M}) = \sigma_{\min}(\mathcal{M}) = 1$, and the observable condition number is

$$\kappa_{\text{obs}}(\mathcal{M}) := \frac{\sigma_{\max}(\mathcal{M})}{\sigma_{\min}(\mathcal{M})} = 1. \tag{34}$$

Dispersion ($\mathcal{D}$). Each block P_k in the dispersion operator $\mathcal{D} = \operatorname{blockdiag}(P_1, \dots, P_{N_\lambda})$ is a permutation matrix. Since permutation matrices are orthogonal, $\mathcal{D}$ itself is an orthogonal operator. Therefore, all singular values of $\mathcal{D}$ are exactly 1, yielding:

$$\kappa_{\text{obs}}(\mathcal{D}) = \frac{\sigma_{\max}(\mathcal{D})}{\sigma_{\min}(\mathcal{D})} = \frac{1}{1} = 1. \tag{35}$$

Spectral integration ($\mathcal{S}$). For the operator $\mathcal{S} = \mathbf{1}_{1\times N_\lambda} \otimes I_n$, which is non-square, we compute the singular values via

$$\mathcal{S}\mathcal{S}^\top = (\mathbf{1}_{1\times N_\lambda}\mathbf{1}_{N_\lambda\times 1}) \otimes I_n = N_\lambda I_n. \tag{36}$$

Hence, all nonzero singular values of $\mathcal{S}$ equal $\sqrt{N_\lambda}$ with multiplicity n. Restricting to the observable subspace (i.e., excluding the zero singular-value directions), the observable condition number is

$$\kappa_{\text{obs}}(\mathcal{S}) := \frac{\sigma_{\max}(\mathcal{S})}{\sigma_{\min}(\mathcal{S})} = 1. \tag{37}$$

Coupled operator ($\mathcal{A} = \mathcal{SDM}$). The singular values of the coupled operator $\mathcal{A}$ can be computed via

$$\mathcal{A}\mathcal{A}^\top = \sum_{k=1}^{N_\lambda} P_k \operatorname{diag}(m) P_k^\top, \tag{38}$$

where each term $P_k \operatorname{diag}(m) P_k^\top$ is a diagonal matrix representing the permuted mask in the k-th spectral band. Denote the diagonal entries of $\mathcal{A}\mathcal{A}^\top$ by

$$g_i := (\mathcal{A}\mathcal{A}^\top)_{ii} = \sum_{k=1}^{N_\lambda} m(\mathcal{P}_k^{-1}(i)), \qquad i = 1, \dots, n, \tag{39}$$

so that g_i counts the number of shifted spectral bands contributing an unmasked sample at output pixel i, i.e., $g_i \in \{0, \dots, N_\lambda\}$.

Let the observable index set be $I_{\text{obs}} = \{i : g_i > 0\}$, then the singular values of $\mathcal{A}$ restricted to the observable subspace are $\{\sqrt{g_i} : i \in I_{\text{obs}}\}$, and the observable condition number is

$$\kappa_{\text{obs}}(\mathcal{A}) = \frac{\max_{i\in I_{\text{obs}}} \sqrt{g_i}}{\min_{i\in I_{\text{obs}}} \sqrt{g_i}} = \sqrt{\frac{\max_{i\in I_{\text{obs}}} g_i}{\min_{i\in I_{\text{obs}}} g_i}}. \tag{40}$$

Since $1 \le g_i \le N_\lambda$ for $i \in I_{\text{obs}}$, a universal worst-case bound follows:

$$\kappa_{\text{obs}}(\mathcal{A}) \le \sqrt{N_\lambda}. \tag{41}$$

This bound is achieved when the mask coverage across shifted bands is highly non-uniform, i.e., $\min_i g_i \ll \max_i g_i$, which justifies the entry $\kappa_{\text{obs}}(\mathcal{A}) = \sqrt{N_\lambda}$ reported in Tab. 1.

A.5 Computational Complexity

To compare the per-iteration computational burden of different CASSI operators, we analyze the cost of solving the regularized normal equations associated with each operator. For an arbitrary forward operator $\mathcal{T}$, consider the linear imaging model

$$\boldsymbol{y} = \mathcal{T}\,\boldsymbol{x} + \boldsymbol{\epsilon}, \tag{42}$$

where $\boldsymbol{y}$ is the measurement vector, $\boldsymbol{x}$ is the unknown signal, and $\boldsymbol{\epsilon}$ denotes additive noise.

Under the standard Tikhonov regularization formula, reconstruction amounts to solving

$$\min_{\boldsymbol{x}} \ \|\boldsymbol{y} - \mathcal{T}\boldsymbol{x}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{x}\|_2^2. \tag{43}$$

Taking the derivative of the objective with respect to $\boldsymbol{x}$ and setting it to zero yields the optimality condition

$$(\mathcal{T}^\top\mathcal{T} + \lambda I)\,\boldsymbol{x} = \mathcal{T}^\top\boldsymbol{y}, \tag{44}$$

which is the regularized normal equation solved—either explicitly or implicitly—in many iterative reconstruction algorithms.

Solving equation 44 typically dominates the computational cost of each iteration, since it requires inverting the matrix $\mathcal{T}^\top\mathcal{T} + \lambda I$. Therefore, characterizing the structure of $\mathcal{T}^\top\mathcal{T}$ for each CASSI operator directly determines its per-iteration computational complexity.

Modulation ($\mathcal{M}$). The modulation operator $\mathcal{M}$ is diagonal, with entries determined by the coded aperture mask. Therefore,

$$\mathcal{M}^\top\mathcal{M} = \mathrm{diag}(m_1^2, \ldots, m_{N_\lambda n}^2) \tag{45}$$

is also diagonal. Including the Tikhonov regularization term, the system matrix becomes

$$\mathcal{M}^\top\mathcal{M} + \lambda I = \mathrm{diag}(m_1^2 + \lambda, \ldots, m_{N_\lambda n}^2 + \lambda), \tag{46}$$

whose inverse (or pseudoinverse when $m_i = 0$) reduces to simple elementwise inversion:

$$(\mathcal{M}^\top\mathcal{M} + \lambda I)^{-1}_{ii} = \frac{1}{m_i^2 + \lambda}.$$

Because no matrix multiplications or factorizations are required, solving the regularized normal equation involving $\mathcal{M}$ has linear complexity $\mathcal{O}(N_\lambda n)$.

Permutation ($\mathcal{D}$). The dispersion operator $\mathcal{D}$ in CASSI is a permutation matrix that reorders spectral channels spatially. Permutation matrices are orthogonal, satisfying

$$\mathcal{D}^\top\mathcal{D} = I, \tag{47}$$

and thus leave the Euclidean norm invariant.

Under Tikhonov regularization, the system matrix becomes

$$\mathcal{D}^\top\mathcal{D} + \lambda I = (1 + \lambda)I, \tag{48}$$

whose inverse is trivially

$$(\mathcal{D}^\top\mathcal{D} + \lambda I)^{-1} = \frac{1}{1 + \lambda} I.$$

Both forward and adjoint operations of $\mathcal{D}$ require only index rearrangements and perform no arithmetic operations; therefore, their computational cost is negligible. Consequently, the solving cost associated with $\mathcal{D}$ is effectively zero, which we denote by "–" in the complexity table.

Spectral integration ($\mathcal{S}$). The spectral integration operator $\mathcal{S} = \mathbf{1}_{1\times N_\lambda} \otimes I_n$ collapses all N_λ spectral channels into a single measurement per spatial pixel. Reconstruction under Tikhonov regularization requires solving the normal equation

$$(\mathcal{S}^\top \mathcal{S} + \lambda I)\, \boldsymbol{x} = \mathcal{S}^\top \boldsymbol{y}. \tag{49}$$

The system naturally decouples across spatial pixels: for each pixel i, the corresponding subsystem is

$$(J_{N_\lambda} + \lambda I_{N_\lambda})\, \boldsymbol{x}_i = y_i\, \mathbf{1}, \tag{50}$$

where J_{N_λ} is the $N_\lambda \times N_\lambda$ all-ones matrix and $\boldsymbol{x}_i \in \mathbb{R}^{N_\lambda}$ represents the spectral values at pixel i. This system admits a closed-form solution:

$$\boldsymbol{x}_i = \frac{y_i}{\lambda + N_\lambda}\, \mathbf{1}. \tag{51}$$

Intuitively, the scalar measurement y_i is distributed across all spectral channels in a weighted manner determined by the regularization parameter λ. As a result, each pixel requires only $\mathcal{O}(N_\lambda)$ operations, leading to a total computational cost of $\mathcal{O}(N_\lambda n)$ for an image with n spatial pixels.

Coupled operator ($\mathcal{A} = \mathcal{S}\mathcal{D}\mathcal{M}$). For the operator $\mathcal{A}$, the regularized normal equation is

$$\boldsymbol{x} = (\mathcal{A}^\top \mathcal{A} + \lambda I)^\dagger \mathcal{A}^\top \boldsymbol{y}, \tag{52}$$

with $\boldsymbol{x} \in \mathbb{R}^{N_\lambda n}$. In general, the Gram matrix $\mathcal{A}^\top \mathcal{A} \in \mathbb{R}^{N_\lambda n \times N_\lambda n}$ is dense due to the block structure

$$\mathcal{A}^\top \mathcal{A} = \mathcal{M}\, \mathcal{S}'^\top \mathcal{S}'\, \mathcal{M}, \tag{53}$$

where $[\mathcal{S}'^\top \mathcal{S}']_{k,\ell} = P_k^\top P_\ell$ and P_k are the permutation blocks of the dispersion operator. While the diagonal blocks satisfy $P_k^\top P_k = I_n$, the off-diagonal blocks $P_k^\top P_\ell$ for $k \neq \ell$ are generally nonzero permutations, introducing dense inter-channel coupling.

As a result, explicitly forming and inverting $\mathcal{A}^\top \mathcal{A}$ with a naive dense solver has worst-case computational complexity $\mathcal{O}((N_\lambda n)^3)$, which justifies the conservative entry in the complexity table.

A.6 Summary of algebraic implications

The above algebraic analysis formalizes several qualitative observations in the main text:

- **Information Retention Ratio:** Coupling the three physical processes into the coupled operator $\mathcal{A}$ reduces the information retention ratio. This reflects a multiplicative loss of information due to both masking and spectral aggregation. Consequently, the coupled operator retains less information than any individual operator, indicating an increased intrinsic difficulty of reconstruction.
- **Null-space dimension:** The coupling of modulation and spectral integration enlarges the kernel of $\mathcal{A}$. An increased null-space dimension corresponds to more unrecoverable degrees of freedom, rendering the inverse problem more ill-posed.
- **Observable condition number:** Operator $\mathcal{A}$ exhibits an increased observable condition number, indicating greater spread in its nonzero singular values. This spectral nonuniformity reduces numerical stability and increases the difficulty of reconstruction.
- **Computational complexity:** While the pseudo-inverse operations of individual operators $\mathcal{M}$, $\mathcal{D}$, and $\mathcal{S}$ can be implemented in linear time $\mathcal{O}(N_\lambda n)$, solving the normal equations associated with the composite operator $\mathcal{A}$ has cubic complexity$\mathcal{O}((N_\lambda n)^3)$. This highlights the higher computational cost of the coupled problem in practical reconstruction.

B Loss Function

The network is trained using the ℓ_1 loss, which measures the discrepancy between the reconstructed hyperspectral image (HSI) and the ground truth:

$$\mathcal{L} = \|x - x^*\|_1, \tag{54}$$

where x denotes the reconstructed HSI and x^* represents the ground truth.

We do not impose additional supervision on the intermediate variables z' and z. Applying explicit supervision to intermediate variables requires additional supervisory signals, which may undermine the fairness of comparisons with standard end-to-end coupled approaches.

C Direction-Cycling GeoT for Multi-Orientation Mamba

To enhance spatial–spectral modeling with limited computational overhead, we introduce a direction-cycling geometric transformation (GeoT) before successive Mamba blocks. Rather than performing multiple directional scans within each block, GeoT deterministically reorients the input features so that successive Mamba blocks traverse the feature tensor along different directions.

Let $\mathbf{F} \in \mathbb{R}^{H\times W\times C}$ denote an input feature tensor. Before the i-th Mamba block, we apply

$$\mathrm{GeoT}_i(\mathbf{F}) = \begin{cases} \mathrm{Flip}(\mathbf{F}), & i \bmod 4 = 0, \\ \mathrm{Transpose}_{HW}(\mathbf{F}), & i \bmod 4 = 1, \\ \mathrm{Transpose}_{HC}(\mathbf{F}), & i \bmod 4 = 2, \\ \mathrm{Transpose}_{WC}(\mathbf{F}), & i \bmod 4 = 3, \end{cases} \tag{55}$$

where $\mathrm{Transpose}_{ab}$ exchanges the corresponding dimensions, and $\mathrm{Flip}(\cdot)$ denotes the predefined spatial flip operation. The transformed feature is then processed by the corresponding Mamba block:

$$\widetilde{\mathbf{F}}_i = \mathcal{M}_i(\mathrm{GeoT}_i(\mathbf{F}_i)), \tag{56}$$

followed by the inverse geometric transformation

$$\mathbf{F}_{i+1} = \mathrm{GeoT}_i^{-1}\left(\widetilde{\mathbf{F}}_i\right), \tag{57}$$

to restore the original feature layout. By cycling the geometric orientation across successive blocks, the same lightweight Mamba operator observes the feature tensor under different traversal orders, thereby enriching spatial–spectral interactions without introducing parallel multi-directional branches. Since GeoT only rearranges tensor elements and contains no learnable parameters, it introduces negligible additional computational overhead.

D Details of Mask-Guided Fusion

The detailed architecture of the proposed mask-guided fusion (MGF) is illustrated in fig. 10. At each scale, the spatially aligned coded mask is first resized to match the feature resolution and encoded by a lightweight convolutional layer. The resulting mask embedding is concatenated with the encoder feature and the corresponding decoder feature, incorporating both physical mask information and the current reconstruction context for gate prediction. The concatenated features are processed by two convolutional layers with a GELU activation, followed by a sigmoid function to generate a feature-wise gating map. The predicted gate multiplicatively modulates the encoder feature, producing mask- and context-aware skip features for subsequent decoder reconstruction. Compared with conventional skip connections that directly forward encoder features, MGF adaptively controls skip-feature propagation using both physical and reconstruction information.

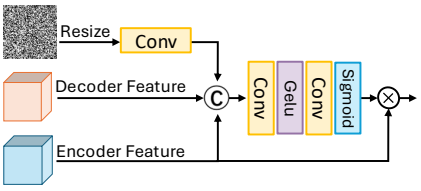


Figure 10: Architecture of the proposed mask-guided fusion (MGF).

E Multi-band Visualization

To further validate the robustness of our approach across the spectral dimension, figs. 11 and 12 present additional visualizations for Scene 6 and Scene 7 over a larger set of wavelength channels. Across all bands, our method consistently preserves high-frequency structures and maintains coherent spectral responses, leading to reconstructions that remain faithful to the ground truth both spatially and spectrally. In comparison, competing methods show band-dependent degradation: fine textures become washed out and noticeable artifacts arise in challenging wavelengths. These multi-band results demonstrate that the proposed reconstruction framework maintains superior performance across full hyperspectral cube, rather than only excelling at a subset of wavelengths.

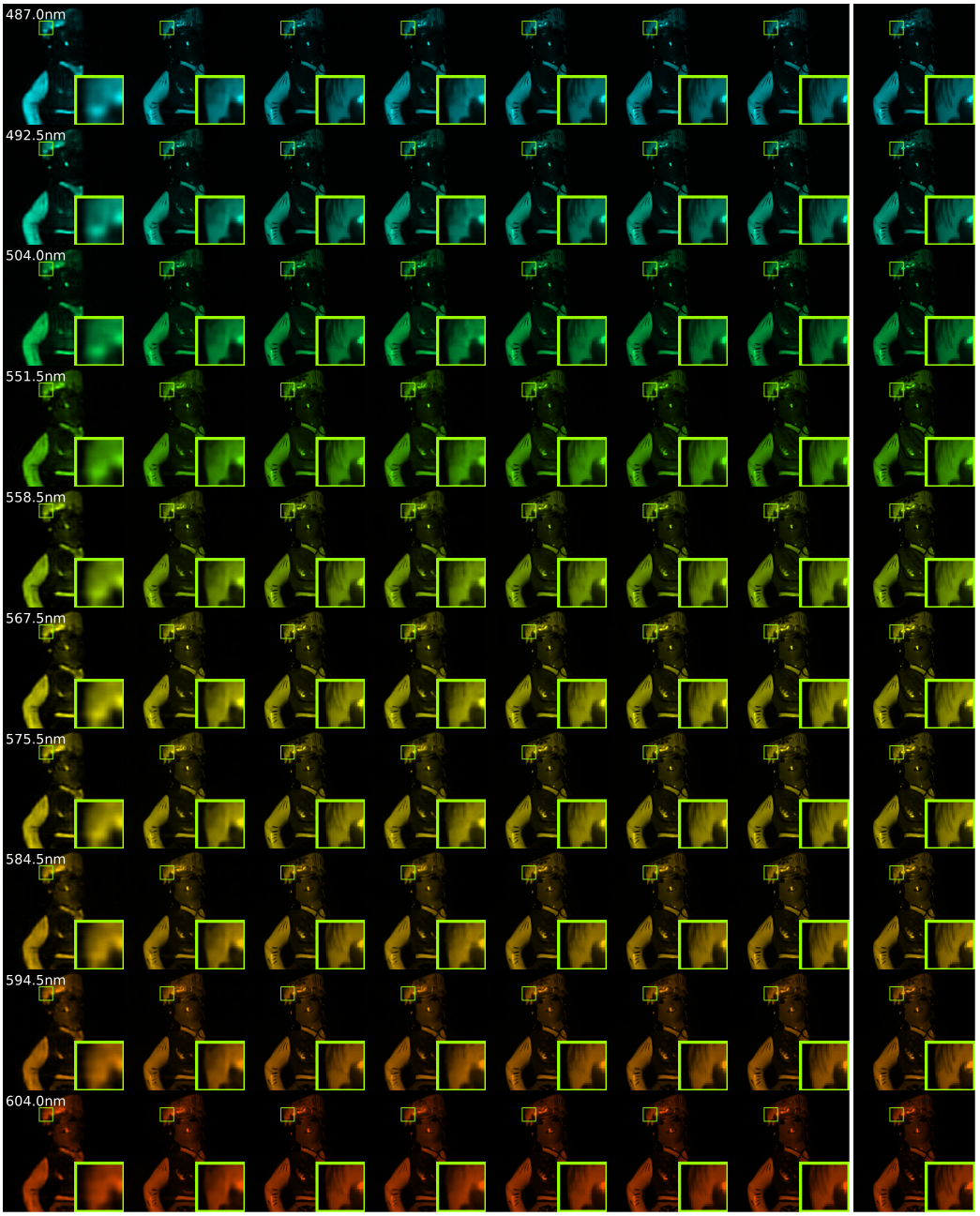


Figure 11: Simulated HSI reconstruction results for **Scene 6** across multiple spectral bands.

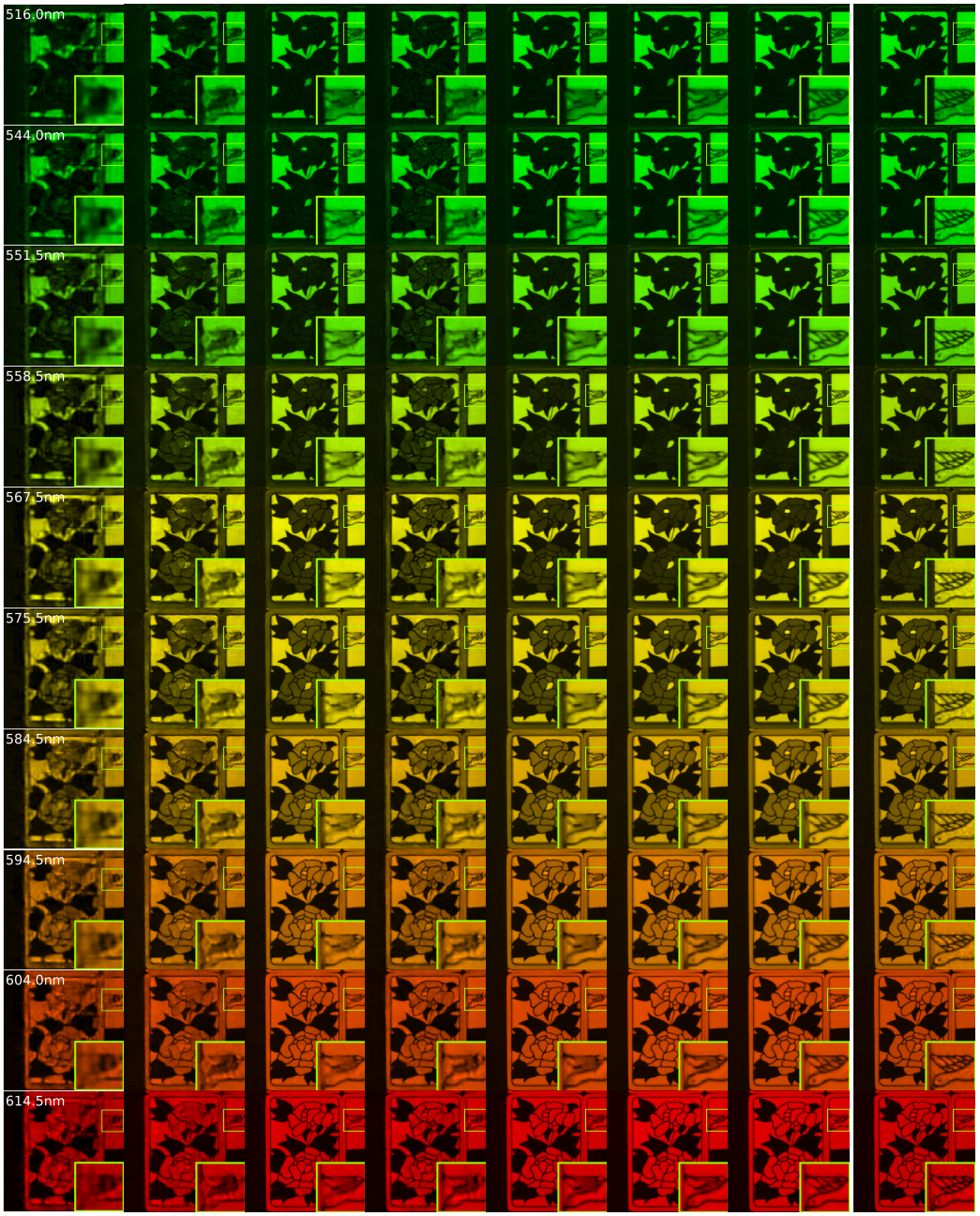


Figure 12: Simulated HSI reconstruction results for **Scene 7** across multiple spectral bands.