Spectral Compressive Imaging via Chromaticity-Intensity Decomposition

Xiaodong Wang [1,2], **Zijun He** [1,2], **Ping Wang** [2],
Lishun Wang [3], **Yanan Hu** [1,2], **Xin Yuan** [2,*]
[1] Zhejiang University,
[2] School of Engineering, Westlake University,
[3] Chengdu Institute of Biology, Chinese Academy of Sciences

Abstract

In coded aperture snapshot spectral imaging (CASSI), the captured measurement entangles spatial and spectral information, posing a severely ill-posed inverse problem for hyperspectral images (HSIs) reconstruction. Moreover, the captured radiance inherently depends on scene illumination, making it difficult to recover the intrinsic spectral reflectance that remains invariant to lighting conditions. To address these challenges, we propose a chromaticity-intensity decomposition framework, which disentangles an HSI into a spatially smooth intensity map and a spectrally variant chromaticity cube. The chromaticity encodes lighting-invariant reflectance, enriched with high-frequency spatial details and local spectral sparsity. Building on this decomposition, we develop CIDNet—a Chromaticity-Intensity Decomposition unfolding network within a dual-camera CASSI system. CIDNet integrates a hybrid spatial-spectral Transformer tailored to reconstruct fine-grained and sparse spectral chromaticity and a degradation-aware, spatially-adaptive noise estimation module that captures anisotropic noise across iterative stages. Extensive experiments on both synthetic and real-world CASSI datasets demonstrate that our method achieves superior performance in both spectral and chromaticity fidelity. Code is released at: `https://github.com/xiaodongwo/CIDNet`.

1 Introduction

Coded aperture snapshot spectral imaging (CASSI) has emerged as a promising architecture for capturing hyperspectral images (HSIs) in a single shot [1, 25, 34]. By jointly modulating the spectral cube with a coded aperture and dispersing it spatially through a prism, CASSI produces a 2D compressed measurement that encodes both spatial and spectral information. This compressive measurement fuses (shears) the spectral bands, making each pixel of the 2D sensor a mixture of many wavelengths. As a result, recovering the full 3D spectral image becomes a severely under-determined, ill-posed inverse problem.

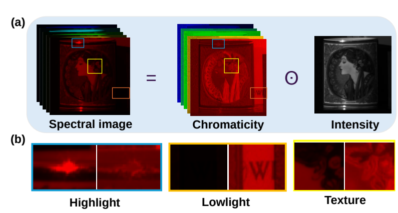

Figure 1: (a) Chrmaticity-Intensity decomposition of HSI images (b) Chromaticity exhibits highlight removal, lowlight enhancement and high-frequency textures.

The difficulty comes mainly from two aspects. One is that **spatial and spectral signals are highly overlapped and entangled in the compressed measurement.** Many works attempt to address this through various priors or deep models, broadly categorized into four paradigms. Optimization-based methods [17, 33] introduce hand-crafted priors such as total variation or low-rank constraints.

*Corresponding Author: Xin Yuan (xyuan@westlake.edu.cn)

However, their performance is often limited in recovering spatial structures, especially under complex textures or noise. Plug-and-Play (PnP) approaches [24, 39] integrate powerful pre-trained denoisers into iterative solvers, yet these methods typically denoise each or few spectral band independently, neglecting spectral correlation and structure. Deep unfolding methods [4, 7, 14, 19, 37, 38] bridge model-based and data-driven paradigms by learning iterative modules guided by the CASSI physics. End-to-end networks [3, 10, 20, 21, 26] leverage CNN or Transformer to directly infer the spectral cube from measurements. These deep learning-based frameworks implicitly exploit spatial-spectral dependencies and have shown promising performance. Recently, diffusion model [22, 30, 35] and Mamba[23] have been used for spectral reconstruction. Nonetheless, all these approaches rely on network backbones to learn spatial-spectral features in an implicit manner. There lacks a clear and interpretable decomposition or quantitative structure that explicitly characterizes the physical roles of spatial and spectral components during reconstruction.

The second challenge is that **existing methods often overlook the impact of illumination.** Since the captured spectral measurement is radiance-based, it inherently entangles the intrinsic surface reflectance with scene illumination. This coupling makes the reconstruction sensitive to lighting variations across time and environments, thereby limiting spectral accuracy. To address similar issues in the RGB image, prior works have explored intrinsic image decomposition [2, 8, 15, 16] and Retinex-based models [28, 29] to explicitly separate reflectance from illumination, enabling applications such as shadow removal and low-light enhancement. In the hyperspectral remote sensing community, several studies have also extended intrinsic decomposition to spectral reflectance and illumination separation [11, 12, 32], offering better invariance to lighting conditions. However, to the best of our knowledge, such decomposition has not yet been incorporated into CASSI reconstruction.

In this paper, we propose a novel chromaticity learning framework for compressive spectral imaging, which leverages a chromaticity-intensity decomposition prior under the CASSI sensing mechanism. Our motivation is illustrated in Fig. 1, where the spectral image cube $\mathbf{X}$ is factorized as:

$$\mathbf{X} = \mathbf{C} \odot \mathbf{I}, \tag{1}$$

where $\mathbf{C}$ denotes the *chromaticity cube* and $\mathbf{I}$ represents the *intensity image*. Notably, $\mathbf{C}$ exhibits several desirable properties: (i) spatially invariant to illumination, suppressing highlights and enhancing details in low-light regions; (ii) spectrally sparse with localized support (as illustrated latter); (iii) enriched with high-frequency texture, essential for fine-detail recovery. In contrast, the intensity component $\mathbf{I}$ captures the global illumination structure in the scene. It is interesting to note that the chromaticity exhibits more intrinsic characteristics of the sample compared to hyperspectral images. Hence, learning the chromaticity instead of HSIs seems to benefit the field more.

Building upon the above observations, we propose a physically interpretable chromaticity-intensity decomposition model tailored for CASSI systems. By leveraging a dual-camera CASSI setup, we validate this decomposition paradigm within both traditional optimization-based solvers and deep unfolding frameworks. To further explore its potential, we design a novel Chromaticity-Intensity Decomposition Network (CIDNet), which incorporates the spectral sparsity of chromaticity through a sparse `TopK` spectral Transformer, and models spatially anisotropic noise via a degradation-aware, spatially-adaptive variance estimator.

In summary, our main contributions are summarized as follows:

i) We propose a novel **chromaticity-intensity decomposition model** for spectral compressive imaging, which explicitly separates hyperspectral images into lighting-invariant chromaticity and smooth intensity. We further validate its effectiveness on optimization-based and unfolding algorithms in a dual-camera CASSI setting.

ii) We develop an **intensity-guided deep unfolding network** that incorporates the chromaticity decomposition into unfolding algorithm. The network features a hybrid spatial-spectral Transformer (HSST) architecture, where the encoder leverages window-based local spatial attention (Spa-LWSA) and the decoder employs sparse `TopK` spectral attention (Spec-TKSA) to capture localized spectral structures.

iii) We introduce a degradation-aware, spatially-adaptive **dual noise estimation module** (DNEM) to model anisotropic noise across different reconstruction stages. This module enables each iteration to adaptively handle varying noise levels across spatial locations.

iv) Extensive experiments on both synthetic and real datasets demonstrate that our method achieves state-of-the-art performance in terms of spectral reconstruction and chromaticity fidelity.

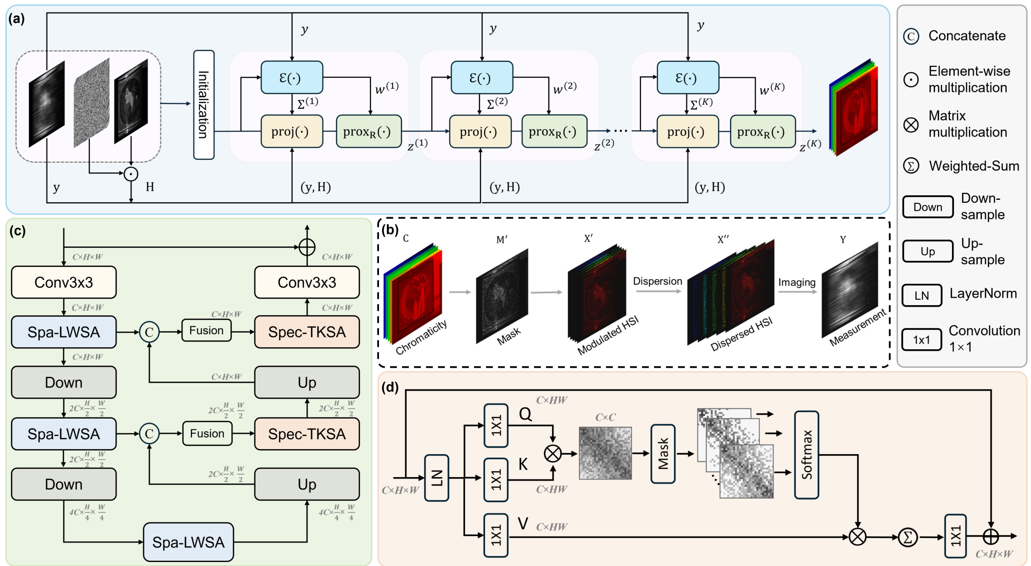

Figure 2: (a) The architecture of our CIDNet with K stages (iterations). (b) The CASSI system uses an intensity-guided mask to modulate the chromaticity. (c) Diagram of asymmetric backbone for our hybrid spatial-spectral Transformer (HSST), with a local window spatial attention (Spa-LWSA) in Encoder and sparse `TopK` spectral attention module (Spec-TKSA) in Decoder. (d) Details of Spec-TKSA.

2 Proposed Method

2.1 Degradation Model of CASSI

Inspired by chromaticity-intensity decomposition in RGB intrinsic image analysis, we extend this concept to the hyperspectral domain. Given a hyperspectral image cube $\mathbf{X} \in \mathbb{R}^{H \times W \times N_\lambda}$, we decompose it into a spatially smooth intensity image $\mathbf{I} \in \mathbb{R}^{H \times W}$ and a chromaticity cube $\mathbf{C} \in \mathbb{R}^{H \times W \times N_\lambda}$ as:

$$\mathbf{X}(u, v, \lambda) = \mathbf{C}(u, v, \lambda) \odot \mathbf{I}(u, v), \tag{2}$$

where (u, v) denotes spatial location, λ is the spectral band index, and $\odot$ denotes pixel-wise multiplication. Specifically, the intensity image is defined as the average spectral energy per pixel:

$$\mathbf{I}(u, v) = \tfrac{1}{N_\lambda} \textstyle\sum_{\lambda=1}^{N_\lambda} \mathbf{X}(u, v, \lambda), \tag{3}$$

We found that intensity image can be approximated as a PAN image in dual-camera CASSI. A detailed proof in this PAN-Intensity Equivalence is provided in supplement materials. Hence, the chromaticity is computed as the normalized spectral signature:

$$\mathbf{C}(u, v, \lambda) = \tfrac{\mathbf{X}(u,v,\lambda)}{\mathbf{I}(u,v)+\epsilon}, \tag{4}$$

where ϵ is a small constant to avoid division by zero. This decomposition separates the multiplicative effect of illumination $\mathbf{I}(u, v)$ from the spectral reflectance $\mathbf{C}(u, v, \lambda)$, which captures intrinsic scene properties. Importantly, $\mathbf{C}$ is invariant to changes in illumination intensity and direction, enabling more robust modeling of reflectance and spectral reconstruction under varying lighting conditions (see supplement materials). After decomposition, the CASSI measurement process can be modeled as follows. The hyperspectral cube $\mathbf{X}$ is modulated by a coded aperture $\mathbf{M} \in \mathbb{R}^{H \times W}$, resulting in a spatially coded cube:

$$\mathbf{X}'(u, v, \lambda) = \mathbf{C}(u, v, \lambda) \odot \mathbf{I}(u, v) \odot \mathbf{M}(u, v). \tag{5}$$

Now we can treat $\mathbf{I}(u, v) \odot \mathbf{M}(u, v)$ as a new formation of coded mask $\mathbf{M}'(u, v) = \mathbf{I}(u, v) \odot \mathbf{M}(u, v)$ incorporating the spatial intensity, we call it intensity-guided mask. This leaves the chromaticity an unknown variable when the intensity is obtained beforehand. Following a typical CASSI formulation, the modulated cube $\mathbf{X}'$ is then passed through a dispersive element that shifts each spectral band λ_{n_λ} by a wavelength-dependent displacement $d(\lambda_{n_\lambda} - \lambda_c)$ along the spatial axis (e.g., the x-axis). The sheared datacube can be expressed as:

$$\mathbf{X}''(u, v, n_\lambda) = \mathbf{X}'(u, v + d(\lambda_{n_\lambda} - \lambda_c), \lambda_{n_\lambda}), \tag{6}$$

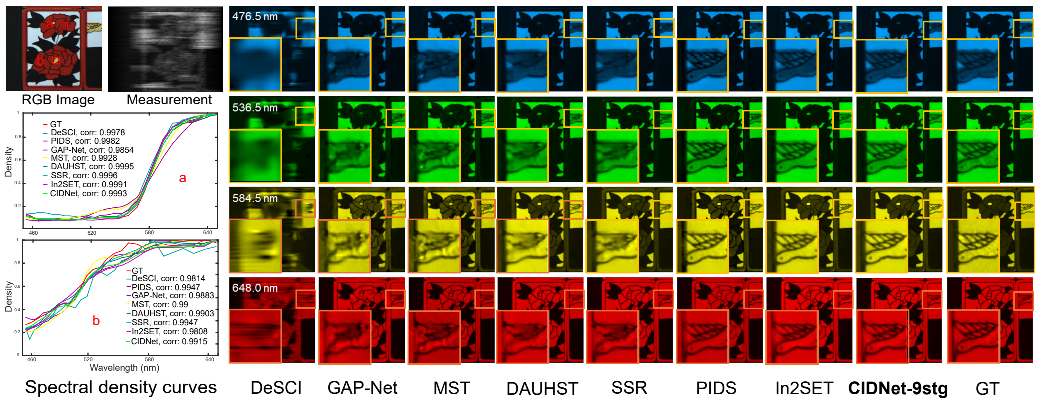

Figure 3: Simulation HSIs reconstruction comparisons of Scene 7 with 4 (out of 28) spectral channels. The left shows the spectral curves corresponding to the two red boxes of the RGB image. The top-right depicts the enlarged patches corresponding to the yellow boxes in the bottom HSIs. Zoom in for a better view.

where λ_c is the reference wavelength that remains unshifted. Finally, the 2D measurement $\mathbf{Y} \in \mathbb{R}^{H\times(W+d\cdot(N_\lambda-1))}$ acquired by the camera is a summation of all dispersed bands:

$$\mathbf{Y}(u,v) = \textstyle\sum_{n_\lambda=1}^{N_\lambda} \mathbf{X}''(u,v,n_\lambda) + \mathbf{N}(u,v), \tag{7}$$

where $\mathbf{N}$ is additive measurement noise. This can be written compactly in vectorized form as:

$$\mathbf{y} = \mathbf{\Phi}(\mathbf{c}\odot\mathbf{i}) + \mathbf{n}, \tag{8}$$

where $\mathbf{c} = \mathrm{vec}(\mathbf{C})$, $\mathbf{i} = \mathrm{vec}(\mathbf{I})$, $\mathbf{\Phi}$ is the sensing matrix determined by the modulation and dispersion process, and $\mathbf{n}$ is the vectorized noise term. Given that the intensity map $\mathbf{I}$ is known (PAN or RGB image in dual-camera CASSI scheme), we formulate the chromaticity-based measurement model as a standard linear inverse problem:

$$\mathbf{y} = \mathbf{H}\mathbf{c} + \mathbf{n}, \tag{9}$$

where $\mathbf{c}$ denotes the vectorized chromaticity, $\mathbf{H}$ is the effective sensing matrix that incorporates both the CASSI modulation-dispersion process and the known intensity modulation, and $\mathbf{n}$ is the vectorized noise.

2.2 Optimization Framework of CASSI

Table 1: Data-consistency projection comparison.

Method	Gradient projection updating
ISTA [36]	$\mathbf{c}^{k+1} = \mathbf{z}^k + \mathbf{H}^\top(\mathbf{y}-\mathbf{H}\mathbf{z}^k)$
GAP [19]	$\mathbf{c}^{k+1} = \mathbf{z}^k + \mathbf{H}^\top(\mathbf{H}\mathbf{H}^\top)^{-1}(\mathbf{y}-\mathbf{H}\mathbf{z}^k)$
HQS [4]	$\mathbf{c}^{k+1} = \mathbf{z}^k + \mathbf{H}^\top(\mathbf{H}\mathbf{H}^\top+\mu\mathbb{I})^{-1}(\mathbf{y}-\mathbf{H}\mathbf{z}^k)$
Ours	$\mathbf{c}^{k+1} = \mathbf{z}^k + \mathbf{H}^\top(\mathbf{H}\mathbf{H}^\top+\mu\mathbf{\Sigma})^{-1}(\mathbf{y}-\mathbf{H}\mathbf{z}^k)$

To characterize realistic imaging noise, we assume an anisotropic Gaussian noise model $\mathbf{n} \sim \mathcal{N}(\mathbf{0},\mathbf{\Sigma})$, where $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1^2,\dots,\sigma_n^2)$ is a diagonal covariance matrix whose diagonal entries σ_i^2 represent the noise variance at the i-th pixel. This implies that the noise is spatially varying but uncorrelated across pixels. Under a Bayesian framework, the posterior probability of the chromaticity $\mathbf{c}$ is given by $p(\mathbf{c}\mid\mathbf{y}) \propto p(\mathbf{y}\mid\mathbf{c})\cdot p(\mathbf{c})$, and the likelihood is:

$$p(\mathbf{y}\mid\mathbf{c},\mathbf{\Sigma}) \propto \exp\left(-\tfrac{1}{2}(\mathbf{y}-\mathbf{H}\mathbf{c})^\top\mathbf{\Sigma}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{c})\right), \tag{10}$$

and $p(\mathbf{c}) \propto \exp(-\tau R(\mathbf{c}))$ is a generic prior over chromaticity with regularization function $R(\cdot)$ and weight $\tau > 0$. Maximizing the posterior leads to the following Maximum A Posteriori (MAP) estimation problem:

$$\hat{\mathbf{c}} = \mathrm{argmin}_{\mathbf{c}}\, \tfrac{1}{2}(\mathbf{y}-\mathbf{H}\mathbf{c})^\top\mathbf{\Sigma}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{c}) + \tau R(\mathbf{c}). \tag{11}$$

When the noise is homoscedastic (i.e., $\sigma_i = 1$), the problem reduces to the common quadratic form: $\hat{\mathbf{c}} = \mathrm{argmin}_{\mathbf{c}}\, \frac{1}{2}\|\mathbf{y}-\mathbf{H}\mathbf{c}\|_2^2 + \tau R(\mathbf{c})$. We rewrite Eq. (11) using an auxiliary variable $\mathbf{z}$:

$$\hat{\mathbf{c}},\hat{\mathbf{z}} = \mathrm{argmin}_{\mathbf{c},\mathbf{z}}\, \tfrac{1}{2}(\mathbf{y}-\mathbf{H}\mathbf{c})^\top\mathbf{\Sigma}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{c}) + \tau R(\mathbf{z}), \quad \text{s.t.} \quad \mathbf{c} = \mathbf{z}. \tag{12}$$

Using the half-quadratic splitting (HQS) framework, Eq. (12) is minimized by solving the following data-consistency and data-prior subproblems iteratively:

$$\mathbf{c}^{(k+1)} = \mathrm{argmin}_{\mathbf{c}}\, \tfrac{1}{2}(\mathbf{y}-\mathbf{H}\mathbf{c})^\top\mathbf{\Sigma}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{c}) + \tfrac{\mu}{2}\|\mathbf{c}-\mathbf{z}^{(k)}\|_2^2, \tag{13}$$

$$\mathbf{z}^{(k+1)} = \mathrm{argmin}_{\mathbf{z}}\, \tfrac{\mu}{2}\|\mathbf{z}-\mathbf{c}^{(k+1)}\|_2^2 + \tau R(\mathbf{z}), \tag{14}$$

where μ is a penalty parameter and k represent kth iteration. The $\mathbf{c}$-subproblem in Eq. (13) is quadratic and has a closed-form solution:

$$\mathbf{c}^{(k+1)} = \left(\mathbf{H}^\top \mathbf{\Sigma}^{-1}\mathbf{H} + \mu\mathbb{I}\right)^{-1} \left(\mathbf{H}^\top \mathbf{\Sigma}^{-1}\mathbf{y} + \mu \mathbf{z}^{(k)}\right), \tag{15}$$

where $\mathbb{I}$ represent identity matrix. Note that $\mathbf{H}^\top \mathbf{\Sigma}^{-1}\mathbf{H}$ is a fat matrix and $(\mathbf{H}^\top \mathbf{\Sigma}^{-1}\mathbf{H} + \mu\mathbb{I})^{-1}$ will be difficult to compute and thus we simplify it based on the Sherman-Morrison-Woodbury formula,

$$\left(\mathbf{H}^\top \mathbf{\Sigma}^{-1}\mathbf{H} + \mu\mathbb{I}\right)^{-1} = \mu^{-1}\mathbb{I} - \mu^{-2}\mathbf{H}^\top \left(\mathbf{\Sigma} + \mu^{-1}\mathbf{H}\mathbf{H}^\top\right)^{-1} \mathbf{H}. \tag{16}$$

In CASSI systems, $\mathbf{H}\mathbf{H}^\top$ is a diagonal matrix defined as $\mathbf{H}\mathbf{H}^\top \triangleq \mathrm{diag}\{h_1, \ldots, h_n\}$. With $\mathbf{\Sigma} \triangleq \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ and detailed derivation in supplement materials, we obtain a generalized form of gradient projection, which is expressed as,

$$\mathbf{c}^{(k+1)} = \mathbf{z}^{(k)} + \mathbf{H}^\top(\mathbf{H}\mathbf{H}^\top + \mu\mathbf{\Sigma})^{-1}(\mathbf{y} - \mathbf{H}\mathbf{z}^{(k)}). \tag{17}$$

Define this gradient projection as $\mathbf{c}^{(k+1)} = \mathrm{proj}_{\mathbf{\Sigma}}(\cdot)$. Interestingly, we observed that this gradient projection resembles previous optimization-based methods but introduces a key change relates to the spatially-varying noise modeling. As summarized in Tab. 1, while traditional ISTA [36], GAP [19], and HQS [4] methods employ static data-consistency steps with fixed regularization, our method proposes a dynamic, spatially-adaptive correction mechanism. Finally, we update $\mathbf{z}^{(k+1)}$ using any proximal operator depending on the prior $R(\cdot)$,

$$\mathbf{z}^{(k+1)} = \mathrm{prox}_{\tau/\mu \cdot R}(\mathbf{c}^{(k+1)}). \tag{18}$$

If the noise variance $\mathbf{\Sigma}$ is known a priori, with Eq. (17) and Eq. (18), this concludes the efficient HQS derivation with anisotropic Gaussian noise. However, in practice, the noise map is unavailable and may vary dynamically across iterations (e.g., in PnP or unfolding methods). To effectively account for the degradation-varying characteristics in the CASSI system, we parameterize the anisotropic noise covariance $\mathbf{\Sigma}^{(k)}$ and the denoising strength τ_k/μ_k in a stage-specific manner. Both parameters are learned by a degradation-aware estimator $\mathcal{E}$ that takes as input the current iterate $\mathbf{z}^{(k)}$ and the measurement $\mathbf{y}$:

$$\{\mathbf{\Sigma}^{(k)}, \tau_k/\mu_k\} = \mathcal{E}(\mathbf{z}^{(k)}, \mathbf{y}). \tag{19}$$

The estimator $\mathcal{E}$ is implemented as a lightweight CNN that jointly captures spatial structure in $\mathbf{z}^{(k)}$ and the encoded degradation in $\mathbf{y}$. A detailed network module is found in supplement materials. We denote $\mathbf{\Sigma}^{(k)}$ and $\omega^{(k)} = \tau^{(k)}/\mu^{(k)}$ as the noise map for gradient projection and proximal mapping (denoiser) respectively (Dual Noise-Estimation Module (DNEM), as we refered to). The estimated $\mathbf{\Sigma}^{(k)}$ reflects the anisotropic uncertainty in the current iterate, and modulates the linear update of $\mathbf{c}^{(k+1)}$ via Eq. (17), while $\omega^{(k)}$ controls the noise level fed into the proximal denoiser for $\mathbf{z}^{(k+1)}$ via Eq. (18). The final iterative process can be expressed as:

$$\begin{aligned}
\{\mathbf{\Sigma}^{(k)}, \omega^{(k)}\} &= \mathcal{E}(\mathbf{z}^{(k)}, \mathbf{y}), \\
\mathbf{c}^{(k+1)} &= \mathrm{proj}_{\mathbf{\Sigma}^{(k)}}(\mathbf{z}^{(k)}) = \mathbf{z}^{(k)} + \mathbf{H}^\top(\mathbf{H}\mathbf{H}^\top + \mathbf{\Sigma}^{(k)})^{-1}(\mathbf{y} - \mathbf{H}\mathbf{z}^{(k)}), \\
\mathbf{z}^{(k+1)} &= \mathrm{prox}_{\omega^{(k)} \cdot R}(\mathbf{c}^{(k+1)}).
\end{aligned} \tag{20}$$

Here, $\mu^{(k)}$ is omitted due to the usage of the network, $R(\cdot)$ is a regularization prior, which could be total variation, or a learned denoiser as in our experiments. This stage-adaptive formulation enables flexible and efficient recovery under spatially variant degradation patterns.

To validate the effectiveness of our chromaticity-intensity decomposition strategy, we explore two integration paradigms: a traditional model-based iterative scheme and a deep unfolding network. The classical iterative algorithm is described in supplement materials, leveraging analytical priors and explicit update rules based on the degradation model. In contrast, our primary design adopts a learnable unfolding structure, as illustrated in Fig. 2. Each stage is composed of a learnable noise estimation, an analytical reconstruction step shown in Eq. (17) and a learned proximal denoiser. This framework offers the interpretability of traditional optimization while benefiting from the expressiveness and efficiency of deep networks.

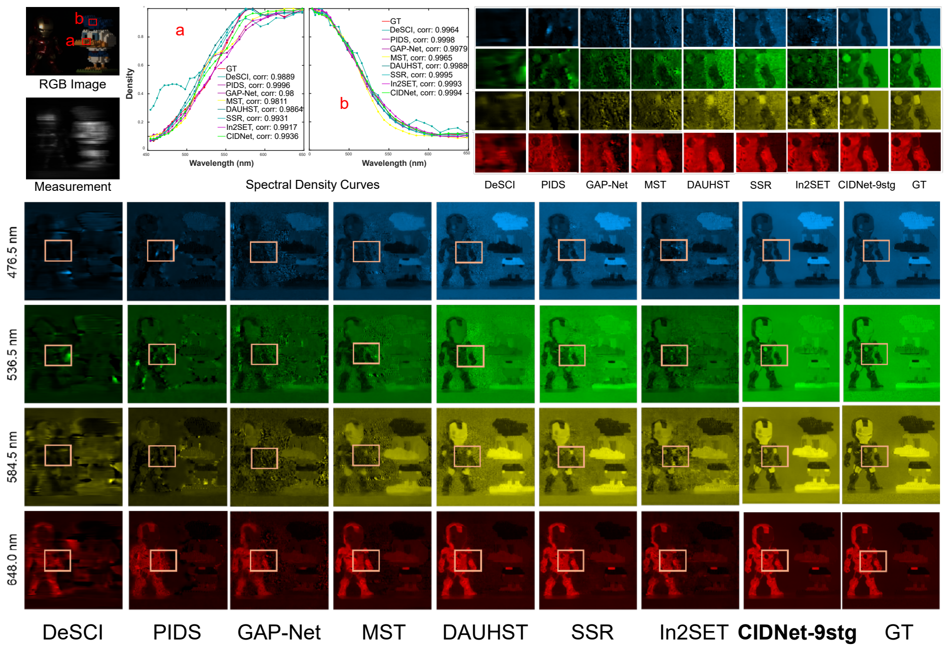

Figure 4: Simulation: **chromaticity** reconstruction of Scene 8 with 4 (out of 28) spectral channels. The spectral curves correspond to the two red boxes in the RGB image (top-middle). The top-right depicts the zoomed patches corresponding to the yellow boxes in the bottom chromaticity.

2.3 Hybrid Spatial-Spectral Transformer

To better reconstruct the chromaticity component $\mathbf{C}$, which inherently contains rich spatial textures and locally correlated spectral patterns (illustrated in Fig. 5), we propose an asymmetric UNet backbone using Hybrid Spatial-Spectral Transformer (HSST). This module is specifically designed to simultaneously learn the high-frequency details in spatial dimensions and sparse-local dependencies in the spectral domain, as shown in Fig. 2.

We adopt a dual-branch design: the spatial attention branch captures intra-image textures through a Swin Transformer in Encoder, while the spectral attention branch is tailored to exploit sparse and locally correlated spectral features using a `TopK` spectral attention mechanism in Decoder. This asymmetric design is inspired by [37] and motivated by experimental verification, where the asymmetric design has better reconstruction results.

Spectral Attention. Unlike the spectral correlation in HSIs, chromaticity spectra features exhibit structured sparsity and localized correlation, see Fig. 5. Motivated by this, we introduce a window-based spectral `TopK` attention mechanism, where attention is applied across the spectral channels within each local spatial window. Specifically, each spectral token attends only to its K most relevant spectral neighbors, enforcing both sparsity and locality.

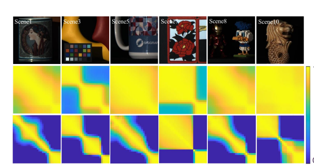

Figure 5: Demonstration of sparse and local spectral correlation of chromaticity. Top: RGB contents of the benchmark testing data. Middle: spectral correlation coefficient matrices of the HSIs (28×28). Bottom: Corresponding matrices by the chromaticity.

Given an input feature cube $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, we first divide it into non-overlapping spatial windows of size $N \times N$, resulting in a batch of local cubes $\{\mathbf{X}_w\} \subset \mathbb{R}^{N^2 \times C}$. Within each window, we perform spectral self-attention across the C channels for every spatial location. We begin by computing the query, key, and value embeddings using learned 1×1 convolutions:

$$\{\mathbf{Q_i}, \mathbf{K_i}, \mathbf{V_i}\} = \mathrm{Conv}_{1\times 1}(\mathbf{X}_w) \in \mathbb{R}^{N^2 \times C \times d}, \tag{21}$$

where d is the embedding dimension per head. To model inter-channel dependencies, we transpose the last two dimensions and perform attention along the channel axis. For each position $i \in \{1, ..., N^2\}$, the attention is computed as:

$$\mathbf{A}_i = \mathrm{Softmax}\left(\mathtt{TopK}\left(\frac{\mathbf{Q}_i \mathbf{K}_i^\top}{\sqrt{d}}\right)\right) \in \mathbb{R}^{C \times C}, \tag{22}$$

Table 2: Comparisons of **HSIs** between CIDNet and SOTA methods on KAIST simulation dataset. PSNR (upper entry in each cell), and SSIM (lower entry in each cell) are reported. The best result is highlighted in bold.

Method	Params(M)	GFlOPs(G)	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg
DeSCI [17]	-	-	27.13 0.748	23.04 0.620	26.62 0.818	34.96 0.897	23.94 0.706	22.38 0.0.683	24.45 0.743	22.03 0.673	24.56 0.732	23.59 0.587	25.27 0.721
GAP-Net [19]	4.27	78.58	33.74 0.911	33.26 0.900	34.28 0.929	41.03 0.967	31.44 0.919	32.40 0.925	32.27 0.902	30.46 0.905	33.51 0.915	30.24 0.895	33.26 0.917
MST-L [3]	2.03	28.15	35.40 0.941	35.87 0.944	36.51 0.953	42.27 0.973	32.77 0.947	34.80 0.955	33.66 0.925	32.67 0.948	35.39 0.949	32.50 0.941	35.18 0.948
DAUHST-9stg [4]	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
SSR-9stg [38]	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 0.980	36.08 0.968	40.69 0.978
PIDS-RGB [5]	-	-	42.09 0.983	40.08 0.949	41.50 0.968	48.55 0.989	40.05 0.982	39.00 0.974	36.63 0.940	37.02 0.948	38.82 0.953	38.64 0.980	40.24 0.967
In2SET-9stg [27]	9.69	59.40	42.56 0.989	46.42 0.994	44.55 **0.986**	**50.63** **0.996**	42.01 0.992	42.49 0.991	41.59 0.983	40.53 0.989	43.83 0.990	**42.33** **0.994**	43.69 0.990
CIDNet-3stg	**1.40**	**24.80**	40.88 0.986	45.39 0.993	43.55 0.983	47.54 0.993	40.37 0.990	41.94 0.901	40.98 0.981	41.11 0.992	42.52 0.987	40.79 0.992	42.51 0.989
CIDNet-5stg	2.33	41.26	41.56 0.987	46.36 0.994	43.98 0.984	47.92 0.993	41.47 0.992	42.27 0.992	41.27 0.982	41.36 0.992	43.90 0.990	40.56 0.992	43.07 0.990
CIDNet-7stg	3.26	57.71	41.66 0.988	46.79 0.995	44.52 0.985	48.51 0.994	41.44 0.992	42.56 0.993	41.46 0.983	41.93 0.993	44.36 0.990	41.49 0.993	43.47 0.990
CIDNet-9stg	4.19	74.16	**42.72** **0.990**	**47.88** **0.996**	**44.87** **0.986**	48.83 0.994	**42.59** **0.993**	**43.01** **0.993**	**42.28** **0.985**	**42.26** **0.994**	**44.68** **0.991**	42.05 **0.994**	**44.12** **0.991**

where $\texttt{TopK}(\cdot)$ retains only the TopK values per row and masks out the rest with $-\infty$ before applying softmax. This yields a sparse attention map across spectral channels for each spatial location i in the window $\mathbf{Z}_i = \mathbf{A}_i\mathbf{V}_i \in \mathbb{R}^{C\times d}$. To improve the robustness and expressiveness of spectral modeling, we further adopt a multi-ratio strategy. Instead of selecting a single sparsity level, we generate multiple attention maps using different TopK ratios (e.g., $\{1/2, 2/3, 3/4, 4/5\}$ of C), and then aggregate them adaptively. Specifically, let $\mathbf{A}^{(r)}$ denote the sparse attention computed under ratio r, and $\alpha^{(r)}$ be a learnable scalar weight. The final attention output is then formulated as:

$$\mathbf{Z}_i = \textstyle\sum_{r\in\mathcal{R}} \alpha^{(r)} \cdot \text{Softmax}(\mathbf{A}_i^{(r)})\mathbf{V}_i, \tag{23}$$

where $\mathcal{R}$ denotes the set of TopK ratios. This fusion across multiple sparsity levels enables the network to capture both dominant and complementary spectral correlations, further improving spectral detail preservation. and reshaped back to the original window structure. After processing all windows, the outputs are stitched to reconstruct the full feature map. This spectral TopK attention not only reduces the computational burden from dense $O(C^2)$ to $O(KC)$, but also explicitly captures the structured sparsity observed in reflectance spectra—where only a few wavelengths contribute significantly. Empirically, we find that this strategy enhances spectral sharpness and suppresses irrelevant cross-band mixing. Following the standard Transformer architecture, we apply a conventional feedforward network after the sparse TopK attention, which is not the focus of our work.

Spatial Attention. Inspired by recent advances in Swin Transformers [18], we employ the Swin Transformer as our spatial modeling backbone. The input feature map $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ is divided into non-overlapping windows of size $M \times M$. Within each window, we perform multi-head self-attention (MSA) by computing

$$\mathbf{Z}_{\texttt{spa}} = \texttt{MSA}\left(\texttt{LN}(\mathbf{X})\right) + \mathbf{X}, \quad \mathbf{Z}_{\text{out}} = \texttt{FFN}\left(\texttt{LN}(\mathbf{Z}_{\texttt{spa}})\right) + \mathbf{Z}_{\texttt{spa}}, \tag{24}$$

where $\texttt{LN}(\cdot)$ denotes layer normalization, and FFN is a standard feedforward network. The relative position bias and shifted-window mechanism in Swin Transformer enhance local texture modeling while maintaining global continuity across windows.

3 Experiments

We conduct comprehensive experiments on both simulated and real-world CASSI systems. The datasets, training settings, and implementation details are introduced as follows.

Table 3: Comparisons of **intensity (left)** and **chromaticity (right)** between CIDNet and SOTA methods on KAIST simulation dataset. PSNR (upper entry in each cell), and SSIM (lower entry in each cell) are reported.

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg
DeSCI [17]	29.42/14.76 0.84/0.49	27.48/23.02 0.72/0.67	31.01/26.75 0.92/0.83	41.71/19.31 0.97/0.75	26.62/20.14 0.81/0.72	25.12/17.66 0.79/0.50	27.08/20.87 0.82/0.60	24.55/19.48 0.77/0.54	29.19/24.90 0.86/0.79	25.61/10.95 0.69/0.32	28.78/19.79 0.82/0.62
GAP-Net [19]	36.78/21.85 0.95/0.62	35.84/20.88 0.93/0.61	39.31/25.66 0.98/0.79	45.90/18.63 0.98/0.66	33.94/23.30 0.94/0.76	34.37/19.43 0.95/0.57	35.65/24.43 0.95/0.64	32.57/20.31 0.94/0.54	37.69/17.34 0.95/0.52	32.05/13.19 0.92/0.39	36.41/20.50 0.95/0.61
MST-L [3]	38.05/23.95 0.96/0.72	38.42/30.26 0.96/0.80	41.07/32.36 0.98/0.89	47.73/19.59 0.99/0.75	35.14/28.89 0.96/0.84	36.44/29.68 0.97/0.78	37.04/27.18 0.96/0.75	34.89/25.94 0.96/0.75	39.26/20.65 0.97/0.74	34.42/17.56 0.95/0.60	38.25/25.61 0.97/0.76
DAUHST-9stg [4]	39.68/25.72 0.97/0.77	40.84/29.54 0.97/0.85	45.08/34.98 0.99/0.94	49.24/31.54 0.99/0.86	37.84/34.08 0.98/0.91	38.79/32.63 0.98/0.84	40.33/31.95 0.98/0.83	36.71/31.39 0.97/0.80	44.55/34.30 0.98/0.90	35.91/26.88 0.96/0.66	40.90/31.30 0.98/0.84
SSR-9stg [38]	41.19/24.45 0.98/0.83	43.35/31.26 0.98/0.93	48.50/40.51 0.99/0.97	50.01/37.65 0.99/0.95	40.85/33.46 0.98/0.95	40.13/36.05 0.98/0.92	41.86/32.72 0.98/0.86	38.44/33.03 0.98/0.91	45.98/37.76 0.99/0.95	37.09/27.86 0.97/0.86	42.74/33.47 0.98/0.91
PIDS [5]	37.44/18.20 0.99/0.72	38.81/17.27 0.98/0.33	34.81/22.74 0.98/0.80	42.83/17.92 0.98/0.68	33.71/19.89 0.98/0.78	36.66/15.90 0.98/0.32	35.95/19.10 0.97/0.61	36.81/17.51 0.98/0.36	37.06/16.24 0.98/0.37	35.43/9.37 0.98/0.15	36.95/17.41 0.98/0.51
In2SET-9stg [27]	58.06/29.42 1.00/0.77	59.61/31.15 1.00/0.83	59.60/36.02 1.00/0.93	62.63/25.56 1.00/0.81	57.55/33.62 1.00/0.90	58.55/27.68 1.00/0.79	57.77/32.05 1.00/0.83	57.82/26.56 1.00/0.76	59.44/25.26 1.00/0.84	56.52/12.93 1.00/0.49	58.75/28.03 1.00/0.80
CIDNet-3stg	63.78/28.02 1.00/0.83	64.70/31.10 1.00/0.93	63.18/39.77 1.00/0.97	**67.12**/36.14 1.00/0.96	61.77/37.54 1.00/0.96	64.73/35.88 1.00/0.93	62.72/32.71 1.00/0.85	65.24/33.97 1.00/0.94	**63.65**/37.60 1.00/0.95	64.93/33.15 1.00/0.89	64.18/34.59 1.00/0.92
CIDNet-5stg	**64.67**/29.70 1.00/0.85	**65.40**/33.41 1.00/0.94	**63.41**/39.54 1.00/0.97	66.83/38.02 1.00/0.96	**64.03**/38.46 1.00/0.96	**65.85**/36.36 1.00/0.93	**63.18**/32.70 1.00/0.86	**65.84/36.70** 1.00/0.94	**63.65**/38.45 1.00/0.96	**64.98**/32.59 1.00/0.89	**64.78**/35.59 1.00/0.93
CIDNet-7stg	61.89/**30.81** 1.00/0.85	62.96/32.06 1.00/0.94	60.99/41.17 1.00/0.97	65.42/38.06 1.00/0.96	61.26/36.91 1.00/0.96	63.87/36.49 1.00/0.93	60.32/33.23 1.00/0.86	63.88/35.98 1.00/0.95	61.14/**38.80** 1.00/0.96	62.52/33.43 1.00/0.89	62.43/35.69 1.00/0.93
CIDNet-9stg	62.28/25.89 1.00/0.86	63.54/**34.29** 1.00/0.95	61.42/**41.48** 1.00/0.97	65.86/**38.22** 1.00/0.96	61.26/**39.71** 1.00/0.97	64.00/**37.10** 1.00/0.94	60.80/**33.31** 1.00/0.87	64.13/36.69 1.00/0.95	62.02/38.00 1.00/0.96	62.71/**33.43** 1.00/0.90	62.80/**35.81** 1.00/0.93

Simulation Dataset. We adopt two widely used hyperspectral datasets: **CAVE** [31] and **KAIST** [6]. The CAVE dataset contains 32 hyperspectral images with a spatial resolution of 512×512. The KAIST dataset provides 30 high-resolution hyperspectral scenes of size 2704×3376. We obtain ground-truth multi-spectral chromaticity and intensity using chromaticity-intensity decomposition Eq. (28). Following prior works [20, 21, 3], we use all CAVE images for training and select 10 scenes from KAIST for evaluation.

Implementation Details. Our model is implemented in PyTorch and trained using the Adam optimizer [13] for 300 epochs. The initial learning rate is set to 4×10^{-4} and updated using a cosine annealing schedule. We employ the ℓ_2 loss between the reconstructed chromaticity and ground-truth chromaticity as the objective function. For training, we randomly extract 3D hyperspectral patches from each scene. For simulated data, the patch size is $256 \times 256 \times 28$, for real-world data, we use patches of size $350 \times 260 \times 26$. In simulation experiments, the forward imaging model is configured with a dispersion shift step $d = 2$, directing dispersion along the horizontal axis (rightward). In real-world scenarios, we assume a vertical dispersion direction and set $d = 1$, consistent with the dual-camera hardware setup. All experiments are conducted in Nvidia A40 GPU.

Comparing Methods. We compared the HSIs and chromaticity reconstruction performance of our CIDNet with other 7 SOTA methods, including DeSCI[17], GAP-Net[19], MST-L[3], DAUHST-9stg[4], SSR-9stg[38] and two dual-camera CASSI algorithm: PIDS[5] and In2SET-9stg[27]. PIDS is compared with RGB image as guidance (in original paper). We evaluate the reconstruction performance from two perspectives: hyperspectral images (HSIs) and chromaticity. Unlike the baseline methods that directly reconstruct HSIs, our method assumes that the intensity is known and focuses on reconstructing the chromaticity. For a fair comparison in the HSI domain, we obtain our reconstructed HSIs by multiplying the recovered chromaticity with the known intensity. Conversely, for chromaticity-level comparison, we perform chromaticity–intensity decomposition on the HSIs reconstructed by the baseline methods to extract their chromaticity components. The reconstruction quality of HSIs is evaluated using peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM).

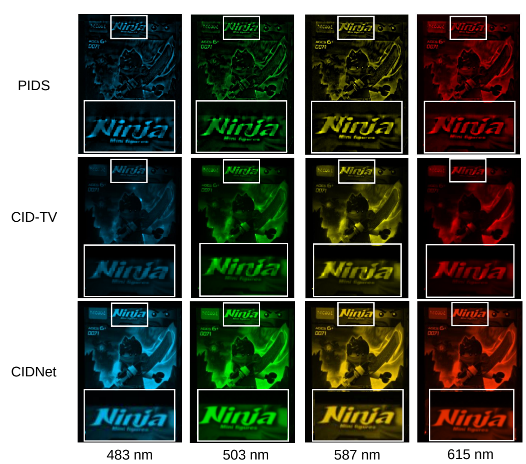

Figure 6: Real-data reconstruction in dual-camera CASSI.

Table 4: Break-down ablation study on individual components of the proposed method.

Base-1	Int.	HSST	DNEM	PSNR	SSIM	Params	FLOPs
✓				35.77	0.949	1.11	16.13
✓	✓			40.83	0.984	1.11	16.13
✓	✓	✓		42.30	0.988	1.44	25.04
✓	✓	✓	✓	**42.51**	**0.989**	1.40	24.80

Table 5: Break-down ablation study on spectral self-attention mechanism.

Method	Base-2	WSSA-WSSA	TKSA-TKSA	LWSA+TKSA	LWSA-TKSA
PSNR	40.95	41.98	42.05	42.36	**42.51**
SSIM	0.985	0.988	0.988	0.988	**0.989**
Params	1.12	1.33	1.33	1.27	1.40
FLOPs	18.03	23.29	25.16	23.47	24.80

Table 6: Break-down ablation study on intensity-guided mask.

Method	ADMM	ADMM-Int	MST	MST-Int	DAUHST	DAUHST-Int
PSNR	24.80	37.09	34.26	37.28	37.21	42.64
SSIM	0.712	0.965	0.935	0.971	0.959	0.989

3.1 Quantitative Results

As shown in Tab. 2, we compare the PSNR and SSIM of HSIs with SOTA methods. our proposed CIDNet excels in 8 out of 10 scenes, particularly in CIDNet-9stg, achieving an average PSNR of 44.12dB and SSIM of 0.991. This significantly surpasses previous unfolding and end-to-end networks, and also in dual-camera CASSI systems, such as In2SET-9stg. A visual comparison is shown in Fig. 3. We provide the visual comparison of simulation Scene7 with 4 out of 28 spectral channels. In addition, we plot the spectral density curves of two regions in the top left RGB image. Our CIDNet-9stg achieves relatively higher spectral accuracy with reference spectra, demonstrating the effectiveness of our method. To verify the reconstructed quality of chromaticity and intensity, we compare the PSNR and SSIM of reconstructed chromaticity and intensity (with decomposition of HSIs), which is shown in Table 3. Note that our method assumes a known intensity. Therefore, to ensure a fair comparison of intensity, we decompose the reconstructed HSIs to extract their intensity component for evaluation. Our CIDNet achieves significant improvements in metrics of chromaticity and intensity. For dual-camera CASSI algorithm PIDS and In2SET, we achieve the best chromaticity metrics with a PSNR of 35.81 and a SSIM of 0.93. A visual comparison of reconstructed chromaticity is shown in Fig. 4. In this research, we used a real-world DCCHI measurement Ninja, taken from publicly available data as detailed in [9]. Fig. 6 illustrates the reconstruction results for four spectral bands in this scene, using two dual-camera CASSI reconstruction algorithms, PIDS and CID-TV, where CID-TV is the iterative CID algorithm using Total Variationa as Regularizer, details can be found in supplement materials. The comparison highlights the superior image restoration quality of our model over other methods, validating its effectiveness and reliability in real-world applications.

3.2 Ablation Study

Effectiveness of Intensity, HSST and DNEM. We verify the effectiveness of our proposed intensity-guided mask and two network module, HSST and DNEM. We adopt Base-1, derived by retaining binary mask and removing spatial-spectral attention and noise estimation from CIDNet-3stg to conduct the ablation study, where we used ground-truth HSIs instead of chromaticity for supervision. Tab. 4 shows the results of PSNR and SSIM of different settings, and our method achieves a significant 6.74dB PSNR improvements compared with Base-1.

Robustness of intensity-guided mask. We test the robustness of intensity-guided mask $\mathbf{M}'$ by employing this intensity mask to ADMM, DAUHST and MST, the result is shown in Tab. 6. We compare the reconstruction quality with/without intensity-guided mask and find that it is significantly improving the base (without intensity) in iterative, end-to-end and unfolding framework. Note that DAUHST achieves slightly higher metric than ours. However its flops and parameters are also greater.

Self-attention scheme comparison. We compare Window-based Spectral Self-Attention (WSSA) [38] and spectral TKSA and its variants within the encoder and decoder. We use $'+'$ to signify a parallel implementation of both attentions (each processing half of the feature channels), and $'-'$ to represent an encoder-decoder implementation. Base-2 is CIDNet-3stg that removes the attention module. Tab. 5 shows the ablation results and LWSA-TKSA yields the most prominent improvement of 1.56dB PSNR compared with Base-2, which shows the effectiveness of our method.

4 Conclusion

We present CIDNet, a novel reconstruction framework for CASSI, which leverages a physically motivated chromaticity-intensity decomposition. By disentangling the hyperspectral image into a spatially smooth intensity map and a spectrally informative chromaticity cube, our method enables lighting-invariant reflectance modeling and better preserves spatial-spectral details. We have designed a hybrid spatial-spectral Transformer to recover the sparse and high-frequency chromaticity components and introduced a degradation-aware unfolding strategy with spatially adaptive noise modeling to handle anisotropic noise inherent in the dual-camera CASSI system. Extensive experiments on both simulated and real-world CASSI datasets validate the effectiveness of our approach, achieving state-of-the-art performance in spectral reconstruction and chromaticity fidelity. This work highlights the benefit of physics-aware decomposition and hybrid attention mechanisms in addressing the ill-posed inverse problem of CASSI reconstruction.

Acknowledgments

This work was supported by National Key R&D Program of China (2024YFF0505603), the National Natural Science Foundation of China (grant number 62271414), Zhejiang Provincial Distinguished Young Scientist Foundation (grant number LR23F010001), Zhejiang 'Pioneer' and 'Leading Goose' R&D Program (grant number 2024SDXHDX0006, 2024C03182), the Key Project of Westlake Institute for Optoelectronics (grant number 2023GD007), the 2023 International Sci-tech Cooperation Projects under the purview of the 'Innovation Yongjiang 2035' Key R&D Program (grant number 2024Z126), National Natural Science Foundation of China (grant number 62501487). (Corresponding author: Xin Yuan.)

References

[1] Gonzalo R Arce, David J Brady, Lawrence Carin, Henry Arguello, and David S Kittle. Compressive coded aperture spectral imaging: An introduction. *IEEE Signal Processing Magazine*, 31(1):105–115, 2013.

[2] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)*, 33(4):1–12, 2014.

[3] Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17502–17511, 2022.

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

[5] Yurong Chen, Yaonan Wang, and Hui Zhang. Prior image guided snapshot compressive spectral imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11096–11107, 2023.

[6] Inchang Choi, Daniel Jeon, Giljoo Nam, Diego Gutiérrez, and Min Kim. High-quality hyperspectral reconstruction using a spectral prior. *ACM Transactions on Graphics*, 36:1–13, 11 2017.

[7] 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*, pages 22262–22271, 2023.

[8] Elena Garces, Carlos Rodriguez-Pardo, Dan Casas, and Jorge Lopez-Moreno. A survey on intrinsic images: Delving deep into lambert and beyond. *International Journal of Computer Vision*, 130(3):836–868, 2022.

[9] Wei He, Naoto Yokoya, and Xin Yuan. Fast hyperspectral image recovery of dual-camera compressive hyperspectral imaging via non-iterative subspace-based fusion. *IEEE Transactions on Image Processing*, 30:7170–7183, 2021.

[10] Xiaowan Hu, Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Hdnet: High-resolution dual-domain learning for spectral compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17542–17551, 2022.

[11] Xudong Jin, Yanfeng Gu, and Tianzhu Liu. Intrinsic image recovery from remote sensing hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):224–238, 2018.

[12] Xudong Kang, Shutao Li, Leyuan Fang, and Jón Atli Benediktsson. Intrinsic image decomposition for feature extraction of hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 53(4):2241–2253, 2014.

[13] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] Miaoyu Li, Ying Fu, Ji Liu, and Yulun Zhang. Pixel adaptive deep unfolding transformer for hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12968, 2023.

[15] Zhengqi Li and Noah Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *Proceedings of the European conference on computer vision (ECCV)*, pages 371–387, 2018.

[16] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9039–9048, 2018.

[17] Yang Liu, Xin Yuan, Jinli Suo, David J Brady, and Qionghai Dai. Rank minimization for snapshot compressive imaging. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):2990–3006, 2018.

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[19] Ziyi Meng, Shirin Jalali, and Xin Yuan. Gap-net for snapshot compressive imaging. *arXiv preprint arXiv:2012.08364*, 2020.

[20] 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*, pages 187–204. Springer, 2020.

[21] Xin Miao, Xin Yuan, Yunchen Pu, and Vassilis Athitsos. l-net: Reconstruct hyperspectral images from a snapshot measurement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4059–4069, 2019.

[22] 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*, pages 25297–25306, 2024.

[23] Mengjie Qin, Yuchao Feng, Zongliang Wu, Yulun Zhang, and Xin Yuan. Detail matters: Mamba-inspired joint unfolding network for snapshot spectral compressive imaging. *arXiv preprint arXiv:2501.01262*, 2025.

[24] Haiquan Qiu, Yao Wang, and Deyu Meng. 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*, pages 9127–9136, 2021.

[25] Ashwin Wagadarikar, Renu John, Rebecca Willett, and David Brady. Single disperser design for coded aperture snapshot spectral imaging. *Applied optics*, 47(10):B44–B51, 2008.

[26] Jiamian Wang, Kunpeng Li, Yulun Zhang, Xin Yuan, and Zhiqiang Tao. S^ 2-transformer for mask-aware hyperspectral image reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[27] Xin Wang, Lizhi Wang, Xiangtian Ma, Maoqing Zhang, Lin Zhu, and Hua Huang. In2set: intra-inter similarity exploiting transformer for dual-camera compressive hyperspectral imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24881–24891, 2024.

[28] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

[29] Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2022.

[30] 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*, pages 164–181. Springer, 2024.

[31] Fumihito Yasuma, Tomoo Mitsunaga, Daisuke Iso, and Shree K Nayar. Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum. *IEEE transactions on image processing*, 19(9):2241–2253, 2010.

[32] Wenbo Yu, Lianru Gao, He Huang, Yi Shen, and Gangxiang Shen. Hi 2 d 2 fnet: Hyperspectral intrinsic image decomposition guided data fusion network for hyperspectral and lidar classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

[33] Xin Yuan. Generalized alternating projection based total variation minimization for compressive sensing. In *2016 IEEE International conference on image processing (ICIP)*, pages 2539–2543. IEEE, 2016.

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

[35] Haijin Zeng, Benteng Sun, Yongyong Chen, Jingyong Su, and Yong Xu. Spectral compressive imaging via unmixing-driven subspace diffusion refinement. In *The Thirteenth International Conference on Learning Representations*.

[36] Jian Zhang and Bernard Ghanem. Ista-net: Interpretable optimization-inspired deep network for image compressive sensing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1828–1837, 2018.

[37] 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*, pages 25742–25752, 2024.

[38] 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*, pages 25817–25826, 2024.

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

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly state that we focus on improving the hyperspectral reconstruction community.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We describe the limitations of our work in Appendix.

   Guidelines

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We give the complete proofs of our theoretical results in Appendix **??**.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental results can be reproduced by the updated source code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Link to GitHub will be available upon release of the paper and the source code is zipped aside in the supplementary material.

   Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The detailed experimental setting is shown in Section 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For all algorithms, each task runs 10 instances with different random seeds.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).

- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The computer resources are shown in Section 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics. The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: We do not foresee any societal impact of our work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not foresee any risk for misuse in our work.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We are under licenses of MIT, BSD, and the Python software foundation.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: answerNA

    Justification: The paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Appendix

In the supplementary material, we provide more details that are not in out main paper:

(a) Multispectral Image Formation in Sec. A.1.

(b) Proof of Illumination-Invariance of Chromaticity in Sec. A.2.

(c) Proof of PAN-Intensity Equivalence in Sec. A.3. We demonstrate the PAN image in dual-camera CASSI is equivalent to the Intensity image.

(d) Closed-form Solution of Data-fidelity Term in Sec. A.4.

(e) Traditional Optimization-based Methods in Sec. A.5. Some results are visualized.

(f) Visual Comparison of HSIs and Chromaticity in Sec. A.6.

(g) Noise Map Visualization of DNEM in Sec. A.7.

(h) Limitation and Broader Impact of Our Work in Sec. A.8

A.1 Multispectral Image Formation

Let (u, v) be spatial coordinates and $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ the wavelength. Assume the scene reflectance is $\mathbf{R}(u, v, \lambda)$(this is not the same as data-prior term), and the illumination spectral power is $L(\lambda)$, spatially uniform. Let $s(\lambda)$ be the spectral response function of the grayscale camera, which converts the spectral radiance to a scalar intensity. Then the intensity recorded by the grayscale camera at (u, v) is:

$$\mathbf{I}(u, v) = \int_{\lambda_{\min}}^{\lambda_{\max}} s(\lambda) \cdot L(\lambda) \cdot \mathbf{R}(u, v, \lambda)\, d\lambda. \tag{25}$$

This is the measurement we observe from the grayscale camera. We define a multispectral distribution:

$$\mathbf{X}(u, v, \lambda) = s(\lambda) \cdot L(\lambda) \cdot \mathbf{R}(u, v, \lambda) \tag{26}$$

Then the multispectral chromaticity function is defined as the normalized form of $X(u, v, \lambda)$ over λ :

$$\mathbf{C}(u, v, \lambda) = \frac{\mathbf{X}(u, v, \lambda)}{\mathbf{I}(u, v)} = \frac{s(\lambda) \cdot L(\lambda) \cdot \mathbf{R}(u, v, \lambda)}{\int s(\lambda) \cdot L(\lambda) \cdot \mathbf{R}(u, v, \lambda)\, d\lambda}, \tag{27}$$

which leads to the final formulation of spectral product of intensity and chromaticity:

$$\mathbf{X}(u, v, \lambda) = \mathbf{C}(u, v, \lambda) \odot \mathbf{I}(u, v). \tag{28}$$

Next we demonstrate that the chromaticity is illumination invariant, and the intensity can be obtained via a dual-camera setting.

A.2 Proof of Illumination-Invariance of Chromaticity

Now suppose the illumination changes globally:

$$L(\lambda) \to \alpha \cdot L(\lambda), \quad \alpha > 0$$

Then:

$$\mathbf{C}'(u, v, \lambda) = \frac{s(\lambda) \cdot \alpha L(\lambda) \cdot \mathbf{R}(u, v, \lambda)}{\int s(\lambda') \cdot \alpha L(\lambda') \cdot \mathbf{R}(u, v, \lambda')\, d\lambda'} \tag{29}$$

$$= \frac{s(\lambda) \cdot L(\lambda) \cdot \mathbf{R}(u, v, \lambda)}{\int s(\lambda') \cdot L(\lambda') \cdot \mathbf{R}(u, v, \lambda')\, d\lambda'} = \mathbf{C}(u, v, \lambda) \tag{30}$$

Thus, the chromaticity $\mathbf{C}(u, v, \lambda)$ is invariant to uniform intensity changes in illumination, even when considering the grayscale camera's spectral sensitivity.

A.3 Proof of PAN-Intensity Equivalence

Proposition A.1 (PAN-Intensity Equivalence Under Uniform Illumination)**.** *In a dual-camera system comprising a CASSI sensor and a grayscale PAN camera exposed under the same illumination $L(\lambda)$, let $s(\lambda)$ denote the spectral response of the camera. The PAN image $\mathbf{I}_{\mathrm{PAN}}(u, v)$ provides a relative estimate of the scene intensity $\mathbf{I}(u, v)$ defined by the chromaticity-intensity decomposition of the hyperspectral image $\mathbf{X}(u, v, \lambda)$. That is,*

$$\mathbf{I}_{\mathrm{PAN}}(u, v) \approx k \cdot \mathbf{I}(u, v), \tag{31}$$

where k is a scalar constant that is approximately invariant across spatial coordinates (u, v).

Proof. Our derivation begins with the Retinex theory, which decomposes an image into chromaticity and intensity components. For hyperspectral images, this decomposition is generalized as:

$$\mathbf{X}(u, v, \lambda) = \mathbf{C}(u, v, \lambda) \cdot \mathbf{I}(u, v), \tag{32}$$

where the chromaticity and intensity are defined as:

$$\mathbf{C}(u, v, \lambda) = \frac{\mathbf{X}(u, v, \lambda)}{\int \mathbf{X}(u, v, \lambda')d\lambda'}, \tag{33}$$

$$\mathbf{I}(u, v) = \int \mathbf{X}(u, v, \lambda)d\lambda. \tag{34}$$

In our dual-camera setup with CASSI and PAN sensors under identical illumination $L(\lambda)$, the PAN image formation is modeled as:

$$\mathbf{I}_{\mathrm{PAN}}(u, v) = \int s(\lambda) \cdot L(\lambda) \cdot \mathbf{X}(u, v, \lambda)d\lambda. \tag{35}$$

Substituting the $\mathbf{X}$ into the product of chromaticity and intensity yields:

$$\mathbf{I}_{\mathrm{PAN}}(u, v) = \int s(\lambda)L(\lambda)[\mathbf{C}(u, v, \lambda) \cdot \mathbf{I}(u, v)]d\lambda. \tag{36}$$

Since $\mathbf{I}(u, v)$ is independent of wavelength λ, it can be factored out of the integral:

$$\mathbf{I}_{\mathrm{PAN}}(u, v) = \mathbf{I}(u, v) \cdot \int s(\lambda)L(\lambda)\mathbf{C}(u, v, \lambda)d\lambda. \tag{37}$$

The key insight is that $\mathbf{C}(u, v, \lambda)$ is normalized ($\int \mathbf{C}(u, v, \lambda')d\lambda' = 1$) and exhibits smooth spectral variation, while $s(\lambda)L(\lambda)$ acts as a broadband low-pass filter. This justifies the approximation:

$$\int s(\lambda)L(\lambda)\mathbf{C}(u, v, \lambda)d\lambda \approx k, \tag{38}$$

where k is a spatial-invariant scalar constant. Therefore, we obtain:

$$\mathbf{I}_{\mathrm{PAN}}(u, v) \approx k \cdot \mathbf{I}(u, v). \tag{39}$$

To address scale ambiguity, we normalize the PAN image to $[0, 1]$ during training and inference, which justifies using PAN as a relative intensity estimate in our chromaticity-intensity framework. $\square$

A.4 Closed-form Solution of Data-fidelity Term

The $\mathbf{c}$-subproblem in Eq. (13) is quadratic and has a closed-form solution:

$$\mathbf{c}^{(k+1)} = \left(\mathbf{H}^\top \mathbf{\Sigma}^{-1}\mathbf{H} + \mu\mathbf{I}\right)^{-1} \left(\mathbf{H}^\top \mathbf{\Sigma}^{-1}\mathbf{y} + \mu\mathbf{z}^{(k)}\right). \tag{40}$$

Note that $\mathbf{H}^\top \mathbf{\Sigma}^{-1}\mathbf{H}$ is a fat matrix and $(\mathbf{H}^\top \mathbf{\Sigma}^{-1}\mathbf{H} + \mu\mathbf{I})^{-1}$ will be difficult to compute and thus we simplify it based on the Sherman-Morrison-Woodbury formula,

$$\left(\mathbf{H}^\top \mathbf{\Sigma}^{-1}\mathbf{H} + \mu\mathbf{I}\right)^{-1} = \mu^{-1}\mathbf{I} - \mu^{-2}\mathbf{H}^\top \left(\mathbf{\Sigma} + \mu^{-1}\mathbf{H}\mathbf{H}^\top\right)^{-1}\mathbf{H}. \tag{41}$$

By plugging Eq. (41) into Eq. (40), we formulate it as

$$\mathbf{c}^{(k+1)} = \frac{\mathbf{H}^\top\mathbf{y} + \mu\mathbf{z}^{(k)}}{\mu} - \frac{\mathbf{H}^\top \left(\mathbf{\Sigma} + \mu^{-1}\mathbf{H}\mathbf{H}^\top\right)^{-1}\mathbf{H}\mathbf{H}^\top\mathbf{y}}{\mu^2} - \frac{\mathbf{H}^\top \left(\mathbf{\Sigma} + \mu^{-1}\mathbf{H}\mathbf{H}^\top\right)^{-1}\mathbf{H}\mathbf{z}^{(k)}}{\mu}. \tag{42}$$

In CASSI systems, $\mathbf{H}\mathbf{H}^\top$ is a diagonal matrix defined as $\mathbf{H}\mathbf{H}^\top \triangleq \mathrm{diag}\{h_1, \ldots, h_n\}$. With $\mathbf{\Sigma} \triangleq \mathrm{diag}(\sigma_1^2, \ldots, \sigma_M^2)$, we obtain:

$$\left(\mathbf{\Sigma} + \mu^{-1}\mathbf{H}\mathbf{H}^\top\right)^{-1} = \mathrm{diag}\left\{\frac{\mu}{\mu\sigma_1^2 + h_1}, \ldots, \frac{\mu}{\mu\sigma_n^2 + h_n}\right\}, \tag{43}$$

$$\left(\mathbf{\Sigma} + \mu^{-1}\mathbf{H}\mathbf{H}^\top\right)^{-1}\mathbf{H}\mathbf{H}^\top = \mathrm{diag}\left\{\frac{\mu h_1}{\mu\sigma_1^2 + h_1}, \ldots, \frac{\mu h_n}{\mu\sigma_n^2 + h_n}\right\}. \tag{44}$$

Let $\mathbf{y} \triangleq [y_1, \ldots, y_n]^\top$ and $[\mathbf{H}\mathbf{c}^{(k)}]_i$ denote the i-th element of $\mathbf{H}\mathbf{c}^{(k)}$. We plug Eq. (43) and Eq. (44) into Eq. (42) as

$$\begin{aligned}
\mathbf{c}^{(k+1)} &= \mu^{-1}\mathbf{H}^\top\mathbf{y} + \mathbf{c}^{(k)} - \mu^{-1}\mathbf{H}^\top \left[\frac{y_1 h_1 + \mu[\mathbf{H}\mathbf{c}^{(k)}]_1}{\mu\sigma_1^2 + h_1}, \ldots, \frac{y_n h_n + \mu[\mathbf{H}\mathbf{c}^{(k)}]_n}{\mu\sigma_n^2 + h_n}\right]^\top & (45)\\
&= \mathbf{c}^{(k)} + \mathbf{H}^\top \left[\frac{y_1 - [\mathbf{H}\mathbf{c}^{(k)}]_1}{\mu\sigma_1^2 + h_1}, \ldots, \frac{y_n - [\mathbf{H}\mathbf{c}^{(k)}]_n}{\mu\sigma_n^2 + h_n}\right]^\top. & (46)
\end{aligned}$$

Generally this is a generalized form of gradient descent, which is expressed as,

$$\mathbf{c}^{(k+1)} = \mathbf{z}^{(k)} + \mathbf{H}^\top(\mathbf{H}\mathbf{H}^\top + \mu\mathbf{\Sigma})^{-1}(y - \mathbf{H}\mathbf{z}^{(k)}) \tag{47}$$

A.5 Traditional Optimization-based Methods

We explore two traditional optimization-based paradigms considering a PAN-guided and RGB-guided intensity respectively, where we formulate the optimization problem using TV prior as,

$$\hat{\mathbf{c}} = \mathrm{argmin}_{\mathbf{c}}\ \tfrac{1}{2}||\mathbf{y} - \mathbf{\Phi}(\mathbf{c} \odot \mathbf{i})||_2^2 + \tau\mathbf{TV}(\mathbf{c}), \tag{48}$$

where $\mathbf{\Phi}$ is the sensing matrix determined by the modulation and dispersion process, $\mathbf{c}$ and $\mathbf{i}$ are chromaticity and intensity respectively, the noise estimation term is omitted since it is hard to be estimated in iterative methods. We consider a dual-camera setting where the second camera could be a grayscale or RGB camera, which satisfy $\mathbf{i} = \mathbf{i}^{\mathbf{PAN}}\ or\ \mathbf{i}^{\mathbf{RGB}}$. In RGB scenarios, the RGB-guided intensity is a three-channel image, which cannot be multiplied directly with Φ due to the channel mismatch. Hence, we interpolate the RGB image to the same spectral channels with corresponding HSIs. Using the HQS framework, Eq. (48) is minimized by solving the following subproblems iteratively by introducing $\mathbf{c} = \mathbf{z}$:

$$\mathbf{c}^{(k+1)} = \mathrm{argmin}_{\mathbf{c}}\ \tfrac{1}{2}||\mathbf{y} - \mathbf{H}\mathbf{c}||_2^2 + \tfrac{\mu}{2}\|\mathbf{c} - \mathbf{z}^{(k)}\|_2^2, \tag{49}$$

$$\mathbf{z}^{(k+1)} = \mathrm{argmin}_{\mathbf{z}}\ \tfrac{\mu}{2}\|\mathbf{z} - \mathbf{c}^{(k+1)}\|_2^2 + \tau\mathbf{TV}(\mathbf{z}), \tag{50}$$

Following previous derivation on Eq.49 and traditional TV denoising term Eq.50, we iterate these two steps to approach its finest solution. The reconstruction is compared with iterative methods such as GAP-TV, DeSCI and PIDS and presented in Fig.7. Scene5 is selected for better visulization purpose.

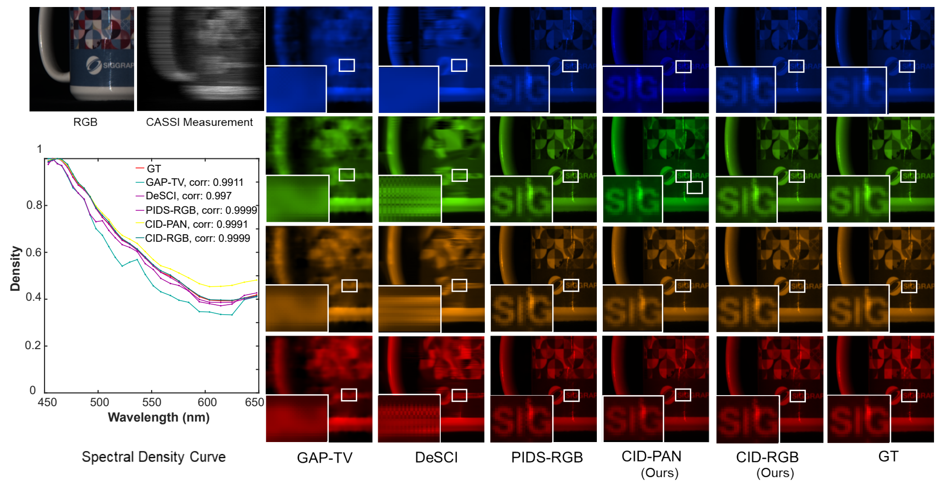

Figure 7: Reconstructed HSIs using traditional optimization-based methods.

A.6 Visual Comparison of HSIs and Chromaticity

We compare the HSIs and chromaticity using KAIST datasets. As shown in Fig. 9, we select 4 out of 10 test datasets, with the top row being the RGB reference and intensity image. Four spectral images are selected for visulization. It can be seen from the figure that chromaticity enhance the low-light regions while preserving more spatial textures as compared with regular spectral images, demonstrating that chromaticity contains more features.

A.7 Noise Map

In this paper we propose a dual noise estimation module for data-fidelity term and denoising network. This network module is designed to estimate a spatially adaptive noise map $\mathcal{E}$, two noise map are output corresponding to the gradient projection noise map and the proximal mapping noise map respectively, where we use convolution and channel attention (CA) to enhances informative channels by modeling inter-channel relationships, as shown in Fig. 8. This structure guarantees positivity and adaptiveness of the output noise map, suitable for uncertainty modeling or variance-aware image restoration tasks. We conduct the experiment and find in CIDNet-3stg, the gradient projection noise map is prominent in the 1st stage and fades away in the latter 2 stages. While the proximal mapping noise map exhibits finer structure in the 2nd and 3rd stages. This is reasonable since more uncertainty is present in the 1st stage introduced by the noisy measurement while the denosing network focus more on the residual learning.

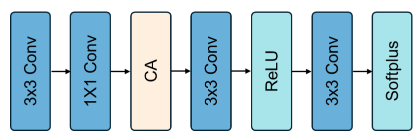

Figure 8: The network structure of step map estimation.

A.8 Limitation and Broader Impact of Our Work

Our work assumes a pre-measured intensity and utilizes a grayscale or RGB image obtained by a dual-camera CASSI system to serve as an intensity image. This is how we obtain intensity image and also our limitation. However, we expect that this decomposition framework is applicable to regular CASSI system. By obtaining the intensity image though a regular CASSI training and then freeze this intensity network, continue training the CIDNet for chromaticity reconstruction. This could be further explored in our future work. Moreover, our chromaticity-intensity decomposition framework opens a new paradigm for low-light or shadow-removal hyper-spectral reconstruction since the chromaticity represents more abundant scene/sample information.

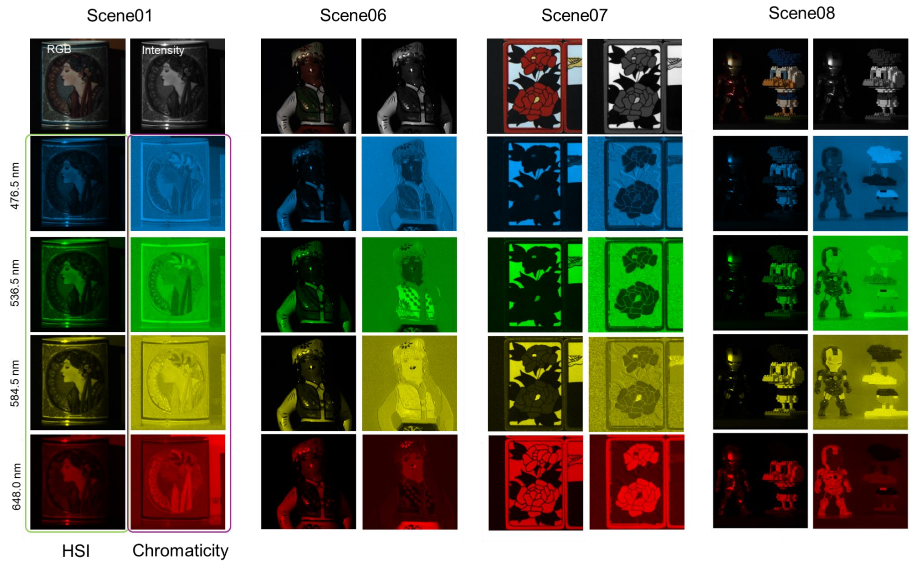

Figure 9: Comparison between HSIs and chromaticity on KAIST test dataset.

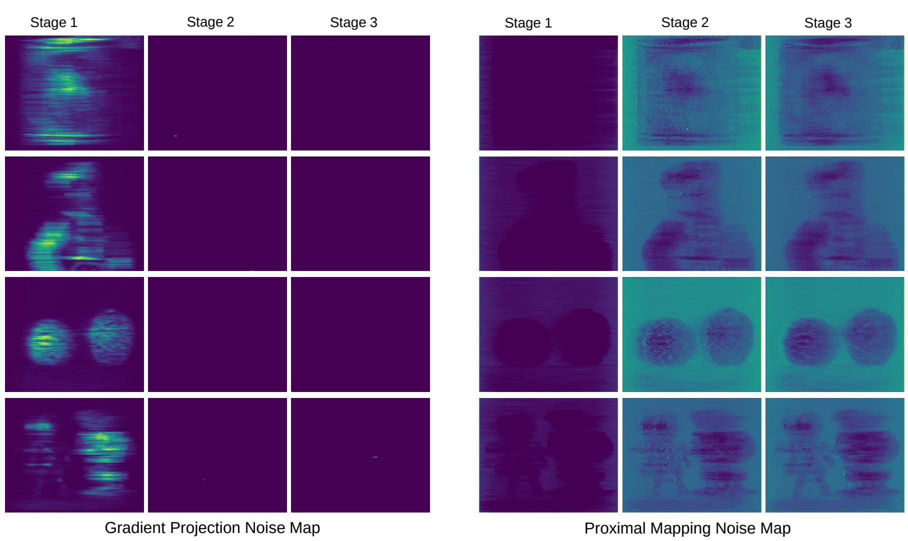

Figure 10: Visulization of dual noise estimation module in CIDNet-3stg.