

From Local to Global: A Progressive Reconstruction Network for Diffractive Snapshot Spectral Imaging

Zhengyue Zhuge¹, Shiqi Chen¹, Chi Zhang¹, Jiahui Xu¹, Tianchen Qiu¹,
Dingchuan Yu¹, and Yueting Chen^{1*}

Zhejiang University
{zhugezy, chenyt}@zju.edu.cn

Abstract. Diffractive Snapshot Spectral Imaging (DSSI) encodes spectral and spatial information into a single 2D measurement via a diffractive optical element (DOE), offering a compact and light-efficient alternative to conventional Coded-Aperture Snapshot Spectral Imaging (CASSI). However, the spatially uneven coding of DSSI, which is mainly concentrated in texture-rich regions, has led existing reconstruction networks to rely heavily on global attention modules to propagate spectral features into texture-less regions, making them short-sighted in locally coded areas and computationally redundant. To address these challenges, we propose Local-to-Global Spectral Reconstruction Network (LGSR-Net), a two-stage progressive framework comprising complementary local and global subnetworks. The local part extracts information from texture-rich regions under direct supervision from effectively coded areas, while the global part models long-range dependencies to facilitate global information flow. We introduce a spatially guided channel attention module to emphasize informative regions and an edge-aware gated fusion mechanism to integrate local features into the global subnetwork’s decoders. Moreover, a coding-aware sample selection strategy and a tailored loss function are designed to improve training efficiency and reconstruction fidelity. Extensive experiments on both simulated and real DSSI datasets show that LGSRNet achieves state-of-the-art reconstruction performance with significantly lower computational cost and latency.

Keywords: Snapshot spectral imaging · Wavefront encoding · Computational imaging · Image reconstruction

1 Introduction

Hyperspectral imaging can capture more wavelength information than a conventional RGB camera, providing rich cues about the material composition and properties of a scene. Hyperspectral images (HSIs) have played an important role in various applications, such as remote sensing [2], agriculture [14], and medical imaging [15]. For fast and compact requirements, several methods attempt to recover spectral information under compressed sensing settings.

* Corresponding author.

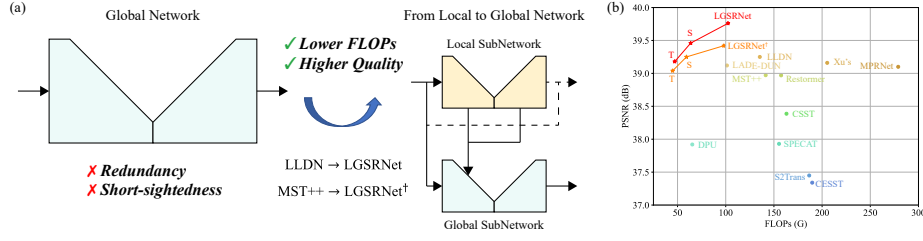


Fig. 1: (a) Motivation and conceptual illustration of the local-to-global network design. Under our local-to-global framework, decomposing a monolithic global network into two smaller subnetworks—a local branch and a global branch—reduces computational complexity while improving reconstruction performance. (b) PSNR vs. FLOPs for different methods on the combination of ICVL [1] and VNIRSR [25] datasets. LGSNet and LGSNet[†] adopt the global modules from LLDN and MST++, respectively, and both achieve competitive performance while significantly reducing FLOPs.

Spectral restoration from RGB images, known as spectral super-resolution (SR) [8,24], aims to reconstruct HSIs from clear RGB images, but the metamerism problem limits reconstruction performance. Coded-Aperture Snapshot Spectral Imaging (CASSI) [9] employs a coded aperture to modulate incoming light, followed by a disperser that spatially shifts different wavelengths onto a 2D sensor. However, such a setting increases system bulkiness and light loss.

Diffraction Snapshot Spectral Imaging (DSSI) [13] is a novel spectral imaging framework that uses a diffractive optical element (DOE) to simultaneously encode spectral and spatial information into a single 2D measurement. Compared to CASSI, DSSI is more compact, lightweight, and exhibits higher light efficiency. Due to the distinct coding mechanism, reconstruction networks designed for CASSI are not directly suitable for DSSI. Moreover, the diffractive coding information is predominantly concentrated in texture-rich regions, making it essential for networks to capture both local and long-range dependencies to recover spectral information across all spatial areas [35].

While networks incorporating global attention modules, such as MST++ [4] and LLDN [35], achieve promising performance, relying solely on global modeling often results in short-sightedness in locally coded regions and redundant computation for local feature extraction. Recent works have explored integrating local and global modeling either alternately [27] or in parallel [20,26] within a single stage. However, these designs remain suboptimal in our experiments. Xu *et al.* [25] proposed a two-stage framework for DSSI that sequentially performs image deblurring and spectral SR, reflecting a similar idea of separating local and global modeling. Nevertheless, the absence of information interaction between stages leads to higher computational cost and limited reconstruction fidelity.

Motivated by Xu’s method [25] and MPRNet [30], we reformulate DSSI reconstruction as a progressive process that follows the spatially uneven DSSI coding order, as shown in Fig. 1(a): one subnetwork first recovers effectively coded local regions, and the other propagates their cues globally. This formulation enables

the local subnetwork to efficiently process texture-rich regions through convolutional operations, while the global subnetwork models long-range dependencies. Moreover, the local subnetwork can be directly supervised by effectively coded regions, providing stronger and more interpretable training signals.

Based on these insights, we propose a two-stage Local-to-Global Spectral Reconstruction Network for DSSI, termed LGSNet, which consists of a local and a global subnetwork collaboratively refining spectral recovery. To focus on texture-rich regions where DSSI encodes information most effectively, a spatially-guided channel attention (SGCA) module is incorporated in the local subnetwork. An edge-aware gated fusion (EAGF) mechanism is then employed to adaptively integrate the local features into the global subnetwork, ensuring effective cross-stage information flow. Additionally, effective coding-aware training sample selection (ECSS) and a tailored loss are used to guide training and enhance overall reconstruction fidelity. Despite using fewer floating point operations (FLOPs) than prior methods, LGSNet achieves state-of-the-art (SOTA) reconstruction performance, as shown in Fig. 1(b).

The main contributions are summarized as follows:

- We propose LGSNet, a two-stage local-to-global reconstruction network for DSSI, combining local and global features to enhance spectral reconstruction performance while reducing computational complexity.
- We design a spatially guided channel attention module to help the local network focus on texture-rich regions, and an edge-aware gated fusion mechanism to integrate local features into the global network.
- We introduce an effective coding-based training sample selection strategy and a tailored loss function to guide training process.
- Experiments on simulated and real datasets demonstrate that LGSNet achieves SOTA reconstruction quality with superior FLOPs and latency.

2 Related Works

2.1 Hyperspectral Reconstruction

HSI reconstruction tasks can be broadly categorized into three groups: spectral SR, CASSI-based, and DSSI-based methods. Motivated by the spectral SR track introduced in the NTIRE challenges, many RGB-to-HSI networks have been proposed [4, 33]. Owing to their compatible input format, these networks can be directly applied to DSSI systems. Modern CASSI reconstruction methods [3, 7] typically take both the coded measurement and the corresponding mask as inputs, leveraging end-to-end deep networks to reconstruct the HSI. However, due to the distinct diffractive imaging mechanism, CASSI-based methods cannot be directly applied to DSSI reconstruction. For DSSI, Jeon et al. [13] unfolded the iterative optimization process via gradient descent to solve the data-fidelity subproblem, while Lv et al. [16] replaced the fidelity term with a learnable Shift-Shuffle Transformer module. More recently, Zhuge et al. [35]

proposed a local-enhanced Mamba-based network to capture long-range dependencies and improve reconstruction accuracy. Despite these advances, current DSSI reconstruction networks often rely heavily on global attention modules, resulting in high computational costs and redundancy.

2.2 Multi-stage Restoration Networks

Multi-stage restoration frameworks have demonstrated strong capabilities in progressively refining image quality across various low-level vision tasks. Representative designs [6, 30, 31] divide images into smaller patches that are restored and subsequently merged in a hierarchical manner. In the frequency domain, Guo et al. [10] and Nehete et al. [17] employed two-stage architectures to sequentially reconstruct amplitude and phase information for rain removal. In DSSI task, Xu et al. [25] adopted a two-stage strategy that first performs deblurring and then spectral SR. These studies demonstrate the effectiveness of multi-stage refinement; however, effective feature interaction between stages for DSSI reconstruction remains underexplored.

2.3 Local-Global Integration Strategies

To exploit both fine-grained spatial details and global spectral correlations, recent HSI reconstruction methods integrate local and global modeling modules either in parallel [20, 26] or alternately [27]. Similar designs have also been utilized in natural image SR [28] and HSI classification [19] tasks. However, the simple combination of local and global pathways within a single block causes information entanglement in DSSI. In contrast, we design a lightweight two-stage framework that first focuses on local texture-rich regions and then refines global spectral dependencies, achieving SOTA performance with minimal computational cost.

3 Method

For an ideal camera, the image I_c for a given color channel $c \in \{r, g, b\}$ of a spectral scene S can be expressed in the following discrete form:

$$I_c[x, y] = \sum_{i=1}^{N_\lambda} S[x, y, i] \cdot \Omega_c[i], \quad (1)$$

where i denotes the i^{th} spectral channel, N_λ is the number of wavelength channels, and Ω_c denotes the camera response function (CRF).

For PSF-encoded computational imaging, the coded image J_c at color channel c can be modeled as:

$$\begin{aligned} \tilde{J}_c[x, y] &= \sum_{k=1}^K S[x, y, k] * P_c[x, y, k], \\ J_c[x, y] &= \tilde{J}_c[x, y] + \eta_c[x, y], \end{aligned} \quad (2)$$

where $*$ denotes convolution, k indexes latent slices, $P_c[x, y, k]$ is the corresponding point spread function (PSF), and η_c is the noise term. In our DSSI setting, k indexes spectral bands, $K = N_\lambda$, and $P_c[x, y, k] = p[x, y, k] \cdot \Omega_c[k]$, where $p[x, y, k]$ is the normalized PSF and $\Omega_c[k]$ is the camera response function. For other PSF-encoded imaging tasks, k can instead index depth, polarization, or other latent states. We reconstruct an estimated spectrum $\hat{S}$ from the coded image J using our LGSRNet.

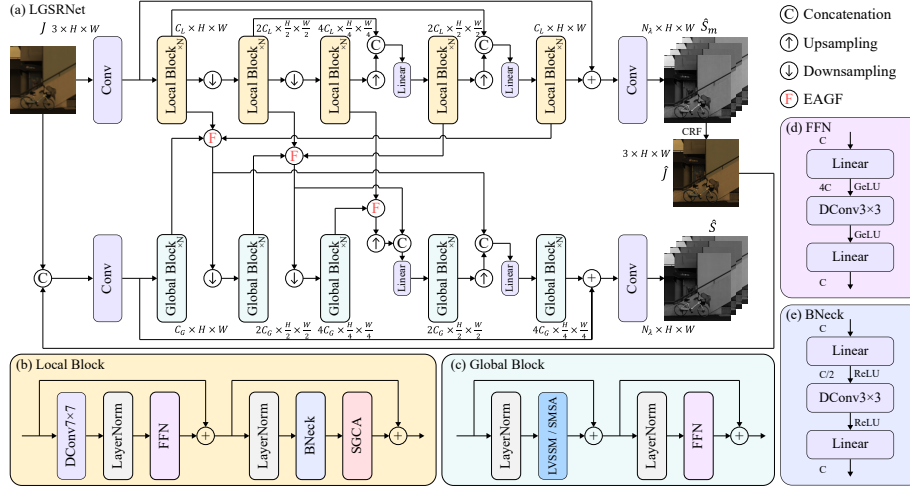


Fig. 2: (a) The overall structure of our LGSRNet. (b) Local block in local subnetwork. (c) Global block in global subnetwork. (d) Bottleneck (BNeck) block. (e) Feed forward network (FFN). SGCA and EAGF modules are detailed in Fig. 3 and 4, respectively.

3.1 Local-to-Global Network

As illustrated in Fig. 2(a), the proposed LGSRNet adopts a two-stage design consisting of a local branch and a global branch, which follow a similar U-shaped encoder-decoder architecture but employ distinct processing blocks to address complementary aspects of DSSI reconstruction.

More generally, this local-to-global principle may be applicable when PSF encoding yields spatially nonuniform effective information, with reliable cues concentrated in texture-rich or high-frequency regions. The local branch first focuses on texture-rich regions where DSSI provides stronger effective coding cues to recover fine-grained spatial details. Its core unit, the local block (see Fig. 2(b)), employs depthwise convolution (DConv) for local feature extraction, followed by a Feed-Forward Network (FFN), shown in Fig. 2(d), to enhance representational capacity. A bottleneck (BNeck) module, shown in Fig. 2(e) further exploits the

low-rank property of hyperspectral data, and the proposed SGCA module leverages spatial saliency to reweight global channel responses. The computation of a local block is defined as:

$$\begin{aligned} F'_{L,l+1} &= FFN(LN(DConv_7(F_{L,l}))) + F_{L,l}, \\ F_{L,l+1} &= SGCA(BNeck(LN(F'_{L,l+1}))) + F'_{L,l+1}. \end{aligned} \quad (3)$$

The local branch produces an intermediate hyperspectral estimate $\hat{S}_m$ that preserves structurally rich spectral details.

The global branch then focuses on long-range dependency modeling to propagate reliable local cues and ensure accurate reconstruction in smooth or weakly coded regions. Its main component, the global block (see Fig. 2(e)), integrates a global attention module (GAM), such as the Local-enhanced Vision State Space Module (LVSSM) [35] or Spectral Multi-head Self-Attention (SMSA) [3], for global context aggregation, followed by FFN for feature expansion and residual connections for stability:

$$\begin{aligned} F'_{G,l+1} &= GAM(LN(F_{G,l})) + F_{G,l}, \\ F_{G,l+1} &= FFN(LN(F'_{G,l+1})) + F'_{G,l+1}. \end{aligned} \quad (4)$$

The two branches are connected through an intermediate RGB bridge and feature fusion (FF) modules. Specifically, the local branch output $\hat{S}_m$ is projected to RGB image $\hat{J}$ via Eq. (1), which serves as a low-rank spectral representation. The concatenation of $\hat{J}$ and the coded image J forms the input to the global branch. During global decoding, EAGF modules integrate features from the local subnetwork into the global branch, adaptively combining fine structural details with global contextual cues. Finally, the global branch generates the refined hyperspectral reconstruction $\hat{S}$.

3.2 Spatially-Guided Channel Attention

The standard channel attention (CA) mechanism [11] and its simplified variant (SCA) [5] aggregate spatial information through global average pooling to reweight feature channels. However, in DSSI encoding, not every spatial location contains valid spectral information.

To address this issue, we propose a Spatially-Guided Channel Attention (SGCA) module that leverages spatial saliency to perform more informative channel-wise modulation, as illustrated in Fig. 3. Given an input feature map F , a Spatial Attention (SA) mechanism [22] is first applied to highlight texture-rich and informative regions, producing a spatially enhanced feature map $SA(F)$, which is then multiplied with the input to form a spatially weighted representation F' . This intermediate feature is subsequently processed by an SCA module to generate adaptive channel weights $SCA(F')$. Finally, these channel weights are applied to the original feature map F , integrating spatially prioritized cues into a global channel reweighting process:

$$F' = SA(F) \cdot F, \quad F'' = SCA(F') \cdot F. \quad (5)$$

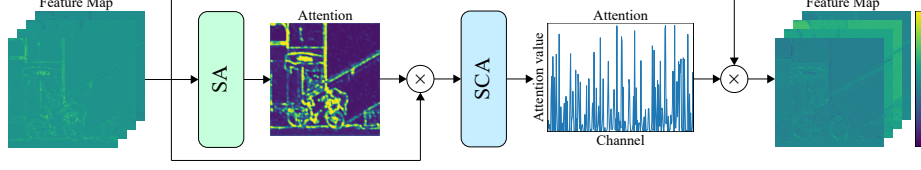


Fig. 3: Spatially-Guided Channel Attention (SGCA) module. Channel attention is derived from spatially weighted feature regions, where spatial attention guides the aggregation of channel-wise responses.

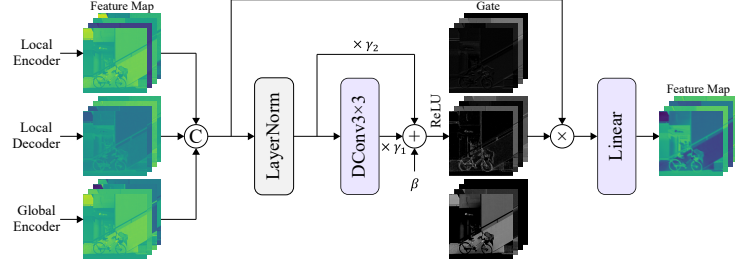


Fig. 4: Edge-Aware Gated Fusion (EAGF) module. LayerNorm and depth-wise convolution are used to emulate edge operators, enabling adaptive edge detection and gated reweighting of multi-source inputs.

As shown in Fig. 3, the SA mechanism assigns higher responses to texture-rich regions that contain effectively encoded information, guiding the subsequent CA module to focus on features derived from these areas.

3.3 Edge-Aware Gated Fusion

For multi-stage networks, MPRNet introduces a Cross-Stage Feature Fusion (CSFF) module [30] to integrate information across different stages. In our local-global two-stage architecture, we propose a more targeted Edge-Aware Gated Fusion (EAGF) module that facilitates adaptive feature integration between the local and global subnetworks of the dual-network framework, as illustrated in Fig. 4. The key motivation stems from the observation that high-frequency (edge-rich) components contain fine structural details, whereas low-frequency components preserve smooth contextual information. By constructing an edge-aware gating map F_{gate} , EAGF adaptively emphasizes or suppresses high-frequency responses, allowing the fusion process to balance edge preservation and contextual smoothness according to local content characteristics.

Given an input feature map F , layer normalization (LN) is applied to mitigate feature-scale inconsistencies:

$$F_{\text{norm}} = \text{LN}(F). \quad (6)$$

A 3×3 DConv is then employed to extract edge information for two complementary branches—an absolute edge branch and a relative edge branch:

$$F_{\text{abs}} = \text{DConv}(F_{\text{norm}}) + \beta_{\text{abs}}, \quad (7)$$

$$F_{\text{rel}} = \text{DConv}(F_{\text{norm}}) + \beta_{\text{rel}} \cdot F_{\text{norm}}, \quad (8)$$

where β_{abs} and γ_{abs} control the absolute edge response, while β_{rel} and γ_{rel} adjust the relative one. For simplicity, the combined gating operation can be expressed as the weighted sum of the two branches followed by activation:

$$\begin{aligned} F_{\text{gate}} &= \text{ReLU}(\gamma_{\text{abs}} F_{\text{abs}} + \gamma_{\text{rel}} F_{\text{rel}}) \\ &= \text{ReLU}(\gamma_1 \text{DConv}(F_{\text{norm}}) + \gamma_2 F_{\text{norm}} + \beta), \end{aligned} \quad (9)$$

where $\gamma_1 = \gamma_{\text{abs}} + \gamma_{\text{rel}}$, $\gamma_2 = \beta_{\text{rel}} \gamma_{\text{rel}}$, and $\beta = \beta_{\text{abs}} \gamma_{\text{abs}}$.

Finally, the gating map F_{gate} modulates the original features via element-wise multiplication, and a linear projection fuses the gated features:

$$F_{\text{fuse}} = \text{Linear}(F_{\text{gate}} \odot F). \quad (10)$$

Through this design, EAGF dynamically balances edge-preserving detail enhancement and context-preserving smoothness, enabling structure-aware and content-adaptive feature fusion between local and global representations.

3.4 Effective Coding-Based Sample Selection

The gradient-based training sample selection strategy proposed in [35] has been shown to improve the training performance. However, it overlooks the impact of noise, particularly Poisson noise, whose magnitude depends on pixel intensity. To address this limitation, we propose an Effective Coding-Based Sample Selection (ECSS) strategy that evaluates whether the informative signal in the coded image J —defined as the difference between the noise-free coded image $\tilde{J}$ and the clean image I —is dominated by noise η .

We first approximate the Poisson noise variance as $\sigma_{P,c}^2 = J_c / 2^{k_c}$, where k_c denotes the number of Poisson noise bits. The total noise variance for channel c is then the sum of Poisson and Gaussian noise variances:

$$\sigma_c^2 = \sigma_{P,c}^2 + \sigma_{G,c}^2 = \frac{\tilde{J}_c}{2^{k_c}} + \sigma_{G,c}^2, \quad (11)$$

where $\sigma_{G,c}^2$ represents the Gaussian noise variance.

Subsequently, we define the effective region $\widetilde{M}_c$ as:

$$\widetilde{M}_c(x, y) = \begin{cases} 1, & |\tilde{J}_c - I_c| > \epsilon \sigma_c \\ 0, & \text{otherwise} \end{cases}, \quad (12)$$

where ϵ is a tunable parameter controlling coding threshold. As illustrated in Fig. 5 (c), the effective regions exist only in certain texture-rich areas, which is

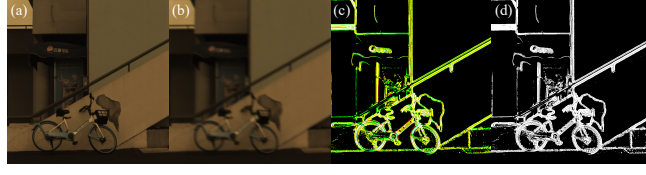


Fig. 5: (a) Clean image I . (b) Coded image J . (c) RGB effective coding region $\widetilde{M}$ and (d) binary mask M ($\epsilon = 6$ in Eq. (12)).

similar to the attention weight map produced by the SA module in Fig. 3. This consistency further validates the rationality of our SGCA.

Finally, a training sample is retained if

$$\sum_{c,x,y} \widetilde{M}_c(x,y) > 100. \quad (13)$$

3.5 Two-Stage Effective Coding-Based Loss

To jointly supervise both stages of LGSNet, we design a two-stage loss that incorporates both full-image and region-aware constraints. For the final output, we adopt an $L1$ +SSIM combination loss following [35]:

$$\mathcal{L}_f(S, \hat{S}) = \frac{\sum(\lambda_1 |S - \hat{S}| + \lambda_2 \text{SSIM}(S, \hat{S}))}{HWN_\lambda}, \quad (14)$$

where λ_1 and λ_2 are empirically set to 0.15 and 0.85, respectively, and H and W denote the spatial dimensions.

For the first-stage output, we constrain supervision to the effectively coded regions. The spatial mask M is computed from $\widetilde{M}$ and extended along the spectral dimension as:

$$M(x, y, i) = \min \left(1, \sum_{c \in \{r,g,b\}} \widetilde{M}_c \right), \quad i = 1, \dots, N_\lambda. \quad (15)$$

The corresponding masked reconstruction loss, termed the Effective Coding-based Loss (ECL), is defined as:

$$\begin{aligned} \mathcal{L}_m(S, \hat{S}_m) = & \frac{1}{\sum M} \sum M (\lambda_1 |S - \hat{S}_m| \\ & + \lambda_2 - \lambda_2 \text{SSIM}(S, \hat{S}_m)). \end{aligned} \quad (16)$$

Finally, the overall training loss combines both terms:

$$\mathcal{L} = \mathcal{L}_f + \mathcal{L}_m. \quad (17)$$

Table 1: The quantitative results on ICVL and VNIRSR datasets.

Method	Params	GFLOPs	ICVL			VNIRSR		
			PSNR $\uparrow$	SAM $\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$	SAM $\downarrow$	SSIM $\uparrow$
UFormer [21]	28.88M	154.62	37.21	0.0646	0.9675	31.95	0.0714	0.9445
CESST [26]	2.92M	189.74	40.77	0.0379	0.9735	35.06	0.0544	0.9502
S^2 Trans [20]	2.67M	186.62	40.85	0.0381	0.9741	35.18	0.0532	0.9484
DPU [32]	3.72M	65.31	41.14	0.0374	0.9772	35.78	0.0506	0.9560
SPECAT [27]	3.54M	155.32	41.33	0.0343	0.9756	35.67	0.0505	0.9516
CSST [16]	11.88M	163.03	41.74	0.0325	0.9768	36.15	0.0472	0.9545
Restormer [29]	26.27M	157.46	42.18	0.0305	0.9780	36.83	0.0438	0.9575
LADE-DUN [23]	6.05M	101.26	42.16	0.0306	0.9767	37.09	0.0434	0.9550
MST++ [4]	10.54M	141.46	42.29	0.0293	0.9781	36.76	0.0446	0.9568
MPRNet [30]	9.05M	278.96	42.19	0.0297	0.9778	37.03	0.0429	0.9572
Xu's [25]	17.34M	205.30	42.54	0.0296	0.9792	36.91	0.0462	0.9562
LLDN [35]	9.48M	135.36	42.40	0.0300	0.9788	37.15	0.0425	0.9597
LGSRNet [†] -T	3.85M	44.96	42.34	0.0297	0.9781	36.84	0.0442	0.9655
LGSRNet [†] -S	4.87M	59.27	42.50	0.0289	0.9787	37.09	0.0431	0.9584
LGSRNet [†]	8.17M	98.04	42.51	0.0288	0.9789	37.36	0.0417	0.9593
LGSRNet-T	3.85M	47.16	42.35	0.0299	0.9788	37.07	0.0434	0.9589
LGSRNet-S	4.87M	63.67	42.55	0.0290	0.9793	37.40	0.0410	0.9604
LGSRNet	8.14M	102.34	42.76	0.0284	0.9800	37.77	0.0394	0.9615

4 Experiments and Analysis

We conduct comprehensive experiments on two HSI datasets—ICVL [1] and VNIRSR [25]—with wavelength ranges from 480 to 900 nm at 10 nm intervals. The training setup follows [35], with the addition of random transpose augmentation to enhance robustness. During testing, we crop the central 896×896 region and apply two fixed metamers ($\alpha \in 0, 1$) together with flip and transpose augmentations. Since the diffractive encoding in our optical system is direction-dependent, these transformations effectively expand the test set and evaluate the robustness of decoding. This procedure increases the test set size by a factor of 16. The injected noise includes 14-bit Poisson noise and Gaussian noise with a standard deviation of 5×10^{-3} .

LGSRNet employs LVSSM in the global blocks, whereas LGSRNet[†] adopts SMSA. All models share the same local subnetwork depth $N_L = 2$. Three model scales are configured as follows:

- Standard: $C_L = 64$, $C_G = 72$, global depth $N_G = 2$;
- Small (S): $C_L = 48$, $C_G = 56$, global depth $N_G = 2$;
- Tiny (T): same channel dimensions as Small, global depth $N_G = 1$.

The effective coding threshold ϵ in the ECSS module is fixed at 6 for all variants. For fair comparison, all learning-based baselines are retrained with the same data, augmentations, noise model, and $L_1 + \text{SSIM}$ objective, with DSSI adaptations detailed in the supplementary material.

4.1 Simulation

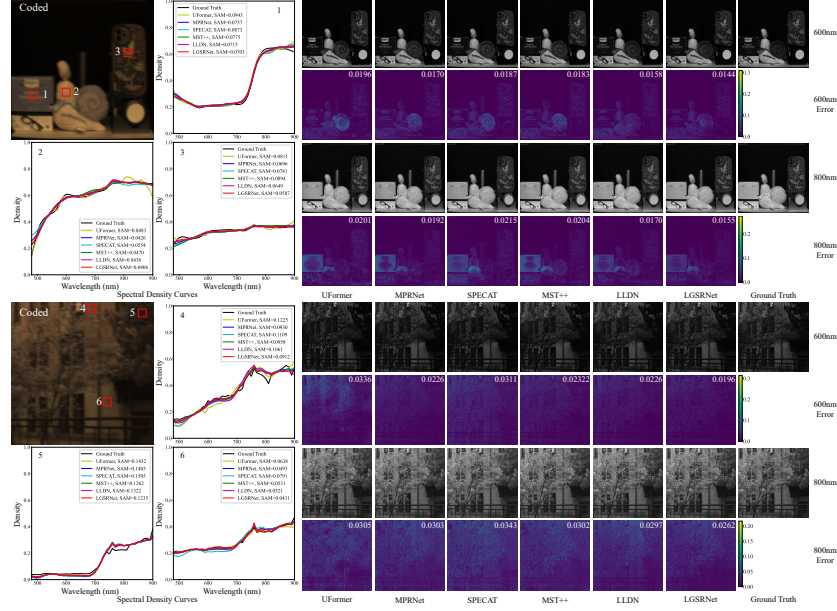


Fig. 6: Visual comparison of reconstructed indoor and outdoor scenes. The spectral density curves are computed from regions marked by red boxes. Error maps with RMSE values are provided in the top-right corner. Please zoom in for better view.

We compare LGSRNet and LGSRNet-T with several SOTA hyperspectral reconstruction models, including spectral SR methods (MST++ [4], CESST [26]), CASSI-based models (SPECAT [27], DPU [32], LADE-DUN [23], S^2 Trans [20]), and DSSI-specific networks (CSST [16], Xu’s [25], LLDN [35]). Among them, DPU, LADE-DUN, and CSST are deep unfolding methods. In addition, we include natural image restoration networks (UFormer [21], MPRNet [30], Restormer [29]) for a comprehensive comparison. Quantitative results, including PSNR, SAM, SSIM, parameter count, and FLOPs, are summarized in Tab. 1, while The PSNR vs. FLOPs results are illustrated in Fig. 1(b).

Our LGSRNet consistently outperforms all compared methods with a lower computational cost. Specifically, on the ICVL dataset, LGSRNet achieves 42.76 dB PSNR, surpassing Xu’s method by 0.22 dB with only half the FLOPs. On the VNIRSR dataset, it reaches 37.77 dB PSNR, exceeding the LLDN by 0.62 dB. On the combination set, LGSRNet-T achieves the second-best performance even with significantly lower FLOPs. In terms of inference latency on a single RTX4090, LLDN requires 172 ms per image, while LGSRNet reduces it to 135 ms, with LGSRNet-S and LGSRNet-T further decreasing the latency to 98 ms and 69 ms, respectively. With the same global module configuration, the

proposed local-to-global architecture substantially improves practical inference speed. Similarly, LGSRNet[†], LGSRNet[†]-S and LGSRNet[†]-T adopt the same SMSA global block as MST++, and they maintain competitive performance while significantly reducing computational cost.

Notably, UFormer exhibits inferior performance due to the absence of a global attention mechanism. CESST, SPECAT, and S^2 Trans yield suboptimal results, owing to their alternating or parallel strategies for integrating local and global modules. These quantitative results validate the effectiveness of our LGSRNet in enhancing both accuracy and efficiency, and our local-to-global two-stage design effectively leverages the interaction between local and global modeling.

Visual comparisons for indoor and outdoor scenes are presented in Fig. 6. LGSRNet achieves the lowest SAM across all six selected regions of interest. In both high-frequency regions (e.g., leaves and the phone shell) and low-frequency regions (e.g., boxes and toys), the error maps of the proposed method appear consistently darker than those of competing approaches, indicating smaller reconstruction errors. This observation is further supported by the RMSE values, which consistently demonstrate lower errors for our method.

4.2 Real-world

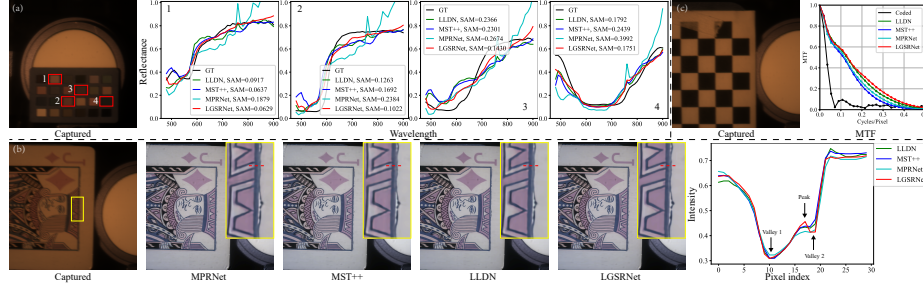


Fig. 7: (a) Real-world color checker and reconstructed spectral reflectance for four color patches highlighted by the red boxes. (b) Real-world playing cards reconstruction results with consistent gain and white balance. The zoomed-in regions correspond to the yellow box and the line intensity profile corresponds to the red dashed line. (c) Real-world checkerboard and reconstructed MTF.

We further validate our method on real-world DSSI data from [35], as shown in Fig. 7. For the color checker, the spectral reflectance of each patch is obtained by normalizing the measured spectra against that of a standard whiteboard. Our method yields reflectance curves that most closely match the GT. For the playing cards, all methods can reproduce the printing imperfections along the card borders, but our reconstruction produces the clearest visual results, with sharper edges and cleaner background restoration. The line profile in Fig. 7(b) further shows that LGSRNet more clearly separates adjacent printed line structures in

this texture-rich region. In particular, the white card background (appearing orange in the figure) is more uniform and clean.

To assess spatial fidelity, the reconstructed HSIs are converted to RGB and then transformed into the L channel of the Lab color space for modulation transfer function (MTF) computation using the slanted-edge method (Fig. 7(c)). Across most spatial frequencies, our method achieves consistently higher MTF values, indicating superior spatial sharpness and fidelity.

4.3 Ablation Study

In this section, we show the results on the combination of ICVL and VNIRS. **Two-stage structure.** We first validate the effectiveness of the proposed local-to-global two-stage design. As shown in Tab. 2, we compare four alternative configurations: (1) a global-only single-stage network, (2) local and global branches operating in parallel, (3) alternately connected local and global modules, and (4) a global-to-global two-stage network. Under comparable numbers of parameters and FLOPs, the parallel or alternating arrangements of local and global modules do not outperform the single-stage global network. In contrast, our local-to-global framework improves the global-only baseline from 39.35 to 39.76 dB PSNR while reducing GFLOPs from 118.07 to 102.34. This confirms the advantage of progressively modeling local and global dependencies for DSSI reconstruction.

Table 2: Ablation study of two-stage design.

Method	Params	GFLOPs	PSNR	SAM
global	8.16M	118.07	39.35	0.0368
local	9.30M	104.16	39.18	0.0375
parallel	8.20M	105.63	38.34	0.0427
alternate	8.29M	106.48	38.92	0.0396
global+global	8.18M	113.81	39.65	0.0357
local+global (ours)	8.14M	102.34	39.76	0.0350

Network modules. To evaluate the contribution of individual modules, we conduct a comprehensive component-wise ablation study, summarized in Fig. 8. Specifically, we either replace our SGCA with the SCA module from NAFNet [5] or completely remove the CA module in the local network, and we either substitute our EAGF with the Cross-Stage Feature Fusion (CSFF) module from MPRNet or remove the FF module between the local and global networks. These variations allow us to isolate the effects of the CA and FF modules on the reconstruction performance.

The results show that both the CA and FF modules play significant roles in improving spectral reconstruction. Notably, without FF, introducing SCA further degrades performance, likely because its simple average-pooling strategy aggregates less informative regions. In contrast, our EAGF alleviates this

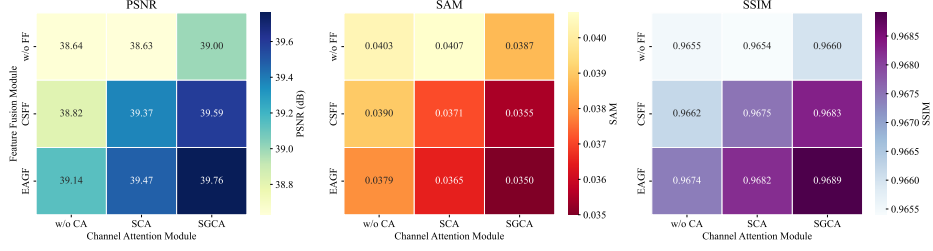


Fig. 8: Ablation heatmap of network components. Combination of different feature fusion (FF) and channel attention (CA) modules. Our proposed EAGF and SGCA modules yield the best performance, and their combination provides a synergistic improvement that exceeds the sum of their individual contributions.

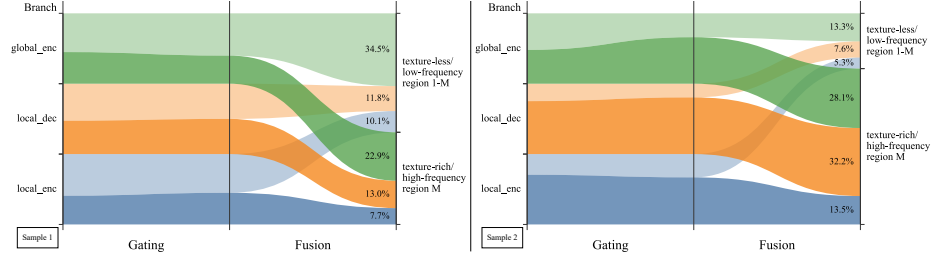


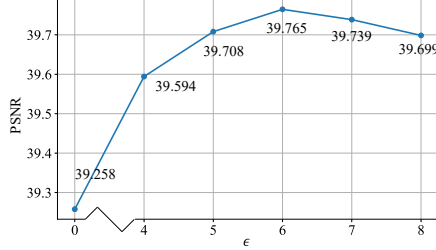
Fig. 9: Energy flow in EAGF. Regional energies are normalized within each gated branch; fusion percentages are computed from the weighted, unnormalized gated-branch energies.

problem and improves PSNR by 0.50 dB under the same setting. Without FF, incorporating SGCA yields a 0.36 dB gain. When combined, the overall improvement reaches 1.12 dB—exceeding the sum of their individual contributions—demonstrating their strong synergistic effect in enhancing spectral reconstruction. These findings highlight that carefully designed attention mechanisms and inter-stage feature fusion are critical for our two-stage design. To further examine how EAGF mediates local-to-global information transfer, Fig. 9 visualizes the regional feature energy in effectively coded (M) and weakly coded ($1 - M$) regions. EAGF regulates both branches before fusion, supporting the propagation of reliable local evidence while preserving contextual information in weakly coded regions. Additional local/global combination ablations are provided in the supplementary material.

Sample selection and loss function. To validate the effectiveness of our proposed strategy, we compare it with two alternatives: training (1) without any sample selection and (2) with the gradient-based selection scheme [35]. They supervise the intermediate output $\hat{S}_m$ using the unmasked loss $\mathcal{L}_f(S, \hat{S}_m)$. Under our proposed ECSS strategy, we further investigate the impact of intermediate supervision and the effectiveness of the proposed ECL $\mathcal{L}_m(S, \hat{S}_m)$. As summarized in Tab. 3, ECSS outperforms both alternatives under the same loss setting.

Table 3: Ablation study of sample selection and loss function.

Selection	Intermediate Loss	PSNR	SAM
$\times$	$\mathcal{L}_f(S, \hat{S}_m)$	39.26	0.0374
gradient	$\mathcal{L}_f(S, \hat{S}_m)$	39.44	0.0364
ECSS	$\times$	38.87	0.0387
ECSS	$\mathcal{L}_f(S, \hat{S}_m)$	39.50	0.0359
ECSS	$\mathcal{L}_m(S, \hat{S}_m)$	39.76	0.0350

**Fig. 10:** PSNR vs. coding threshold parameter ϵ .

Moreover, applying intermediate supervision to $\hat{S}_m$ improves reconstruction accuracy, while adopting the ECL provides additional performance gains.

Effective coding threshold. We further examine the influence of the effective coding threshold ϵ in our sample selection strategy. As ϵ increases, the selected samples become more informative yet fewer in number, resulting in a non-monotonic trend in PSNR that peaks at $\epsilon = 6$ (see Fig. 10).

5 Conclusion

Based on the spatially uneven diffractive encoding characteristics of DSSI, we design LGSRNet, a two-stage local-to-global spectral reconstruction network. By progressively modeling fine-grained local textures and long-range global dependencies, LGSRNet effectively extracts spectral information and mitigates the nonuniform spatial encoding inherent in diffractive imaging with less computation redundancy. The proposed SGCA and EAGF modules enhance region-aware representation and local-to-global fusion, while the coding-aware sample selection and loss function further improve training efficiency and reconstruction fidelity. Extensive experiments on simulated and real DSSI datasets demonstrate that LGSRNet achieves SOTA reconstruction quality with lower computational cost. Beyond HSI reconstruction, this framework may apply to other PSF-encoded computational imaging tasks, such as depth estimation [12, 34], and polarization imaging [18], when effective cues are spatially nonuniform and require local-to-global propagation.

References

1. Arad, B., Ben-Shahar, O.: Sparse recovery of hyperspectral signal from natural rgb images. In: *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII* 14. pp. 19–34. Springer (2016)
2. Borengasser, M., Hungate, W.S., Watkins, R.: *Hyperspectral remote sensing: principles and applications*. CRC press (2007)
3. Cai, Y., Lin, J., Hu, X., Wang, H., Yuan, X., Zhang, Y., Timofte, R., Van Gool, L.: Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 17502–17511 (2022)
4. Cai, Y., Lin, J., Lin, Z., Wang, H., Zhang, Y., Pfister, H., Timofte, R., Van Gool, L.: 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 (2022)
5. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: *European conference on computer vision*. pp. 17–33. Springer (2022)
6. Chen, L., Lu, X., Zhang, J., Chu, X., Chen, C.: Hinet: Half instance normalization network for image restoration. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 182–192 (2021)
7. Chen, Y., Lai, W., He, W., Zhao, X.L., Zeng, J.: Hyperspectral compressive snapshot reconstruction via coupled low-rank subspace representation and self-supervised deep network. *IEEE Transactions on Image Processing* **33**, 926–941 (2024)
8. Galliani, S., Lanaras, C., Marmanis, D., Baltasvias, E., Schindler, K.: Learned spectral super-resolution. *arXiv preprint arXiv:1703.09470* (2017)
9. Gehm, M.E., John, R., Brady, D.J., Willett, R.M., Schulz, T.J.: Single-shot compressive spectral imaging with a dual-disperser architecture. *Optics express* **15**(21), 14013–14027 (2007)
10. Guo, X., Fu, X., Zhou, M., Huang, Z., Peng, J., Zha, Z.J.: Exploring fourier prior for single image rain removal. In: *IJCAI*. pp. 935–941 (2022)
11. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 7132–7141 (2018)
12. Ikoma, H., Nguyen, C.M., Metzler, C.A., Peng, Y., Wetzstein, G.: Depth from defocus with learned optics for imaging and occlusion-aware depth estimation. In: *2021 IEEE International Conference on Computational Photography (ICCP)*. pp. 1–12. IEEE (2021)
13. Jeon, D.S., Baek, S.H., Yi, S., Fu, Q., Dun, X., Heidrich, W., Kim, M.H.: Compact snapshot hyperspectral imaging with diffracted rotation. *ACM Transactions on Graphics (TOG)* **38**(4), 1–13 (2019)
14. Lu, B., Dao, P.D., Liu, J., He, Y., Shang, J.: Recent advances of hyperspectral imaging technology and applications in agriculture. *Remote Sensing* **12**(16), 2659 (2020)
15. Lu, G., Fei, B.: Medical hyperspectral imaging: a review. *Journal of biomedical optics* **19**(1), 010901–010901 (2014)
16. Lv, T., Ye, H., Yuan, Q., Shi, Z., Wang, Y., Wang, S., Cao, X.: Aperture diffraction for compact snapshot spectral imaging. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 10574–10584 (2023)

17. Nehete, H., Monga, A., Kaushik, P., Kaushik, B.K.: Fourier prior-based two-stage architecture for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6014–6023 (2024)
18. Shen, Z., Zhao, F., Jin, C., Wang, S., Cao, L., Yang, Y.: Monocular metasurface camera for passive single-shot 4d imaging. *Nature Communications* **14**(1), 1035 (2023)
19. Sheng, J., Zhou, J., Wang, J., Ye, P., Fan, J.: Dualmamba: A lightweight spectral-spatial mamba-convolution network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* (2024)
20. Wang, J., Li, K., Zhang, Y., Yuan, X., Tao, Z.: s^2 -transformer for mask-aware hyperspectral image reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025)
21. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A general u-shaped transformer for image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17683–17693 (2022)
22. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
23. Wu, Z., Lu, R., Fu, Y., Yuan, X.: Latent diffusion prior enhanced deep unfolding for snapshot spectral compressive imaging. In: European Conference on Computer Vision. pp. 164–181. Springer (2024)
24. Xiong, Z., Shi, Z., Li, H., Wang, L., Liu, D., Wu, F.: Hscnn: Cnn-based hyperspectral image recovery from spectrally undersampled projections. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 518–525 (2017)
25. Xu, H., Zhuge, Z., Chen, B., Xu, N., Xu, J., Zhang, G., He, J., Xu, Z., Li, Q., Chen, Y.: Two-stage framework for reconstructing spectral images from diffraction-blurred images. *Optics and Lasers in Engineering* **186**, 108789 (2025)
26. Yang, X., Chen, J., Yang, Z.: Hyperspectral image reconstruction via combinatorial embedding of cross-channel spatio-spectral clues. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6567–6575 (2024)
27. Yao, Z., Liu, S., Yuan, X., Fang, L.: Specat: Spatial-spectral cumulative-attention transformer for high-resolution hyperspectral image reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25368–25377 (2024)
28. Yu, D., Li, Q., Wang, X., Zhang, Z., Qian, Y., Xu, C.: Dstrans: Dual-stream transformer for hyperspectral image restoration. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3739–3749 (2023)
29. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5728–5739 (2022)
30. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14821–14831 (2021)
31. Zhang, H., Dai, Y., Li, H., Koniusz, P.: Deep stacked hierarchical multi-patch network for image deblurring. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5978–5986 (2019)
32. Zhang, J., Zeng, H., Cao, J., Chen, Y., Yu, D., Zhao, Y.P.: Dual prior unfolding for snapshot compressive imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25742–25752 (2024)

33. Zhao, Y., Po, L.M., Yan, Q., Liu, W., Lin, T.: Hierarchical regression network for spectral reconstruction from rgb images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 422–423 (2020)
34. Zhuge, Z., Xu, H., Chen, S., Xu, J., Li, Q., Feng, H., Xu, Z., Chen, Y.: Calibration-free deep optics for depth estimation with precise simulation. *Optics and Lasers in Engineering* **180**, 108313 (2024)
35. Zhuge, Z., Zhang, C., Xu, J., Li, Q., Chen, S., Chen, Y.: Diffractive snapshot spectral imaging needs long-range dependency. *Optics & Laser Technology* **192**, 113638 (2025)