SPECAT: SPatial-spEctral Cumulative-Attention Transformer for High-Resolution Hyperspectral Image Reconstruction

Zhiyang Yao* Shuyang Liu* Xiaoyun Yuan Lu Fang†
Department of Electronic Engineering, Tsinghua University

Abstract

Compressive spectral image reconstruction is a critical method for acquiring images with high spatial and spectral resolution. Current advanced methods, which involve designing deeper networks or adding more self-attention modules, are limited by the scope of attention modules and the irrelevance of attentions across different dimensions. This leads to difficulties in capturing non-local mutation features in the spatial-spectral domain and results in a significant parameter increase but only limited performance improvement. To address these issues, we propose SPECAT, a SPatial-spEctral Cumulative-Attention Transformer designed for high-resolution hyperspectral image reconstruction. SPECAT utilizes Cumulative-Attention Blocks (CABs) within an efficient hierarchical framework to extract features from non-local spatial-spectral details. Furthermore, it employs a projection-object Dual-domain Loss Function (DLF) to integrate the optical path constraint, a physical aspect often overlooked in current methodologies. Ultimately, SPECAT not only significantly enhances the reconstruction quality of spectral details but also breaks through the bottleneck of mutual restriction between the cost and accuracy in existing algorithms. Our experimental results demonstrate the superiority of SPECAT, achieving 40.3 dB in hyperspectral reconstruction benchmarks, outperforming the state-of-the-art (SOTA) algorithms by 1.2 dB while using only 5% of the network parameters and 10% of the computational cost. The code is available at https://github.com/THU-luvision/SPECAT.

1. Introduction

Improving imaging resolution and increasing the dimensionality of acquired information are critical challenges currently faced in the field of visual imaging [1–4]. Exploiting their unique spectral characteristics, hyperspectral imaging (HSI) play a crucial role in various fields such as precision agriculture [5], defense [6], environmental monitoring [7], and space exploration [8]. Additionally, they significantly contribute to the advancement of computer vision applications, including object tracking [9, 10], material classification [11], feature identification [12, 13], and medical imaging [14]. Traditional imaging methods, which sequentially scan spatial or spectral dimensions and require multiple exposures to reconstruct scene spectral data [15], perform poorly in real-time applications. The development of Compressed Sensing (CS) theory has led to the creation of Snapshot Compressed Imaging (SCI) systems [16], integrating spectral snapshot information into a single 2D metric.

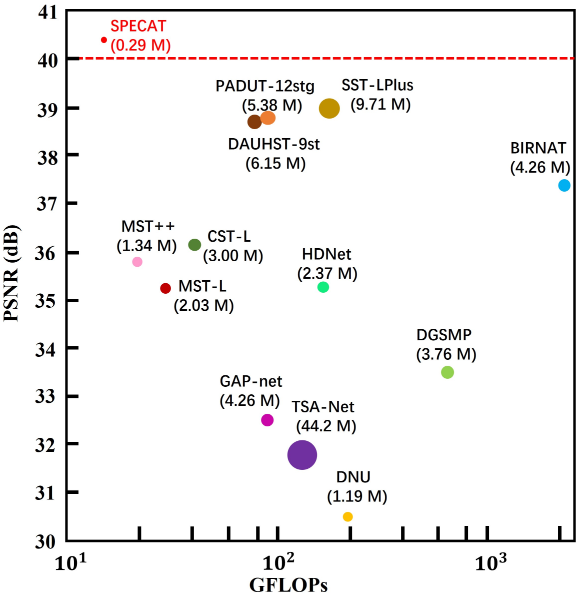


Figure 1. Comparison of PSNR, Params, and GFLOPs between our SPECAT and SOTA HSI reconstruction methods. The vertical axis represents PSNR (dB), the horizontal axis corresponds to GFLOPs (computational cost), and the circle radius signifies Params (memory cost).

Among these systems, Coded Aperture Snapshot Spectral Imaging (CASSI) [17, 18] has emerged as a leading area of research. CASSI system stores three-dimensional hyperspectral images on a two-dimensional sensor by en-

*Equal Contribution , †Corresponding Author.

coding both the spatial and spectral domains, then obtains the complete hyperspectral data from these compressed data by compressive spectral image reconstruction algorithm. Traditional hyperspectral reconstruction methods [19–24] achieve hyperspectral image reconstruction through gradual iterations by utilizing the CS theory and sparse prior. However, their iterative nature and computational intensity often result in slower reconstruction speeds.

Deep learning has been used in compressive spectral reconstruction and has shown significant advantages in high-resolution and rapid hyperspectral image reconstruction. The most effective performers are primarily End-to-End (E2E) deep learning algorithms and Deep Unfolding (DU) networks. E2E algorithms directly learn the mapping from compressive images to their hyperspectral counterparts, offering advantages such as rapid processing and effective performance [25]. Current E2E algorithms extract features at different scales using deeper network structures in a single independent domain [26, 27] and adopt separate spatial sparsity learning and spectral recovery modules [28]. The deeper network structures increase the network complexity and computational demands significantly. The separate learning of the spatial and spectral information faces challenges in capturing global spatial-spectral mutation features.DU networks, which employ network-based sparse prior instead of regularization as in conventional reconstruction algorithms, iteratively restore hyperspectral images based on CS physical models [29, 30]. However, they often have larger parameters and high computational overhead. In short, the current methods encounter a bottleneck in the reconstruction accuracy of spectral details and computational cost.

To address the aforementioned issues, we propose a SPatial-spEctral Cumulative-Attention Transformer (SPECAT) for high-resolution hyperspectral image reconstruction. SPECAT adopts joint spatial and spectral attention through spatial-spectral cumulative-attention blocks (CAB), efficiently extracting non-local mutation features and improving the reconstruction quality of spectral details. By employing multi-head self-attention in the spatial domain, we achieve a sparse representation of image characteristics across different spectral bands. Subsequently, we apply multi-head spectral-wise attention to the spatial features of different spectral bands through a hierarchical structure, capturing subtle features that represent high fidelity and spectral domain mutations. Relative to current methodologies, CAB offers advantages in two aspects. First, cumulative spatial and spectral attention enables global key feature extraction at a single scale. This approach greatly reduces the network depth and computational costs needed for detailed feature extraction. Second, lower-level spatial attention capitalizes on spatial sparsity and highlights mutation areas. This supports upper-level spectral attention in accurately reconstructing mutation features and similarities, thereby increasing the precision of reconstruction.

The main contributions of this paper can be summarized in the following three aspects.

(1) We propose a new method, SPECAT, for high-resolution hyperspectral image reconstruction. Compared with previous state-of-the-art (SOTA) methods, SPECAT achieves, for the first time, a breakthrough of an average 40 dB PSNR with a much smaller number of parameters and computational cost in the simulated hyperspectral reconstruction benchmark with a single camera.

(2) We present a novel attention block, CAB, to efficiently extract non-local mutation features and restore the global crucial information of hyperspectral images.

(3) We study the performance of existing algorithms and our proposed algorithm on the optical filters-based hyperspectral system (e.g. liquid crystal tunable filter-based HSI[31], metasurface HSI[32], Fabry-Pérot filters-based HSI[33]), offering a reliable and effective reconstruction algorithm suitable for future on-chip HSI systems.

2. Related works

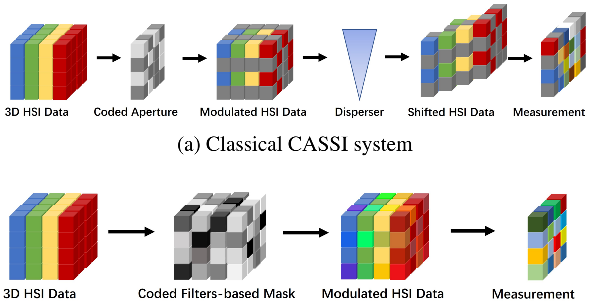
2.1. Hyperspectral Imaging Systems

Since its emergence, HSI has advanced remarkably, diversifying into several forms such as pushbroom [34], whiskbroom[35], and snapshot modalities [36]. Particularly, encoding-based snapshot HSI utilizes a mask for compressed image acquisition in the spatial-spectral domain, offering high temporal, spatial, and spectral resolution. As Fig. 2 shows, classical CASSI systems encode spatial and spectral domains separately, resulting in larger sizes and reduced tolerance to jolts or impacts. Filter-based HSI systems use broadband filters to encode both domains with a single mask. Not only do they offer high light throughput and high resolution, but they also greatly simplify the optical path of the imaging system, allowing for integration onto a chip [31–33, 37]. Therefore, exploring hyperspectral reconstruction in optical filter-based HSI systems is crucial for advancing the miniaturization of spectrometers [38].

2.2. Hyperspectral Reconstruction Algorithm

Traditional hyperspectral reconstruction methods [19–24] are centered around model-driven techniques, mainly relying on handcrafted priors. Due to their limited representational capabilities, they often lack efficiency and flexibility. In contrast, deep learning methods can establish powerful mappings, directly reconstructing hyperspectral images from measurements or aiding the iterative process to generate more optimal prior, significantly enhancing the efficiency and accuracy of hyperspectral image reconstruction. Existing hyperspectral reconstruction methods include algorithms directly targeting a single fixed mask system and those integrating mask optimization into a holistic optimization approach. Although the latter significantly improves reconstruction performance, it increases the complexity and size of the system at the hardware level[39, 40]. Therefore, considering the future miniaturization demands of spectrometers, the former has more advantages.

Among existing methods for CASSI systems with a fixed mask, CNN-based networks [26, 41, 42] excel in extracting local spatial correlations for data reconstruction and offer the advantage of rapid inference. However, they may overlook global features. The development of Transformer[43]-based networks in recent years provides a reliable means for hyperspectral imaging [25, 27–30]. However, to improve the reconstruction accuracy of subtle features, existing methods either separate spatial sparsity learning and spectral recovery modules to reduce redundant information and enhance local attention [28], or use deeper network structures to extract features at different scales through multiple convolutions and downsampling [25, 27, 28, 44]. However, limited by the scope of attention modules and the irrelevance of attention across different dimensions, it is difficult to capture non-local abrupt features in the spatial-spectral domain. Although the DU networks[29, 30] incorporates physical models to improve the reconstruction quality of non-local features to a certain extent, the lack of joint application of spatial-spectral attention also results in limited improvement in detailed reconstruction. This ultimately leads to current SOTA strategies significantly increasing network complexity and computational demands, with only limited performance improvements.



(a) Classical CASSI system

(b) Optical filters-based HSI system

Figure 2. The diagram of the classical CASSI system and optical filters-based HSI system.

3. Methods

3.1. Model of Snapshot HSI System

The concise CASSI system schematic is shown in Fig. 2. Using $\mathbf{X}\in \mathbb{R}^{H\times W\times N_\lambda}$ to represent the three-dimensional HSI data, and setting the projection matrices for spatial encoding and spectral encoding as $\mathbf{M}\in \mathbb{R}^{H\times(W+d(N_\lambda-1))\times N_\lambda}$ (which is generated by shifting the actual physical mask $\mathbf{m}\in \mathbb{R}^{H\times W}$ with d), the two-dimensional data $\mathbf{Yc}\in \mathbb{R}^{H\times(W+d(N_\lambda-1))}$ received by the sensor satisfies Eq. (1). $d\in\mathbb{R}^1$ denotes the shifting step.

$$\mathbf{Yc} = \sum_{\lambda} \mathbf{Mc} \odot \text{shift}(\mathbf{X}) + \mathbf{Gc} \tag{1}$$

Where the shift$(\cdot)$ function characterizes the process in the CASSI system, $\odot$ denotes the element-wise multiplication. After the polychromatic light passes through the dispersive element, a two-dimensional mask $\mathbf{m}$ equivalently generates a three-dimensional mask $\mathbf{M}$, completing the joint spatial spectral domain encoding process, and generating compressed two-dimensional images from the shifted three-dimensional hyperspectral images. And $\mathbf{Gc}\in \mathbb{R}^{H\times(W+d(N_\lambda-1))}$ was the measurement noises.

For a hyperspectral imaging system based on optical filters, a single pixel of a two-dimensional mask can encode all spectral bands (i.e., different spectral bands will have different transmittance rates for that pixel). For a specific spectral band of interest, the transmittance through different mask pixels can be achieved by changing the structure of the filters. In short, by designing different types of filters and mapping them one-to-one with the pixels of the two-dimensional sensor, the spatial-spectral encoding and information compression of the three-dimensional HSI data cube can be accomplished. Eq. (2) provides the spectral compressive imaging model for optical filter-based systems.

$$\mathbf{Y} = \sum_{\lambda} \mathbf{M} \odot \mathbf{X} + \mathbf{G} = \sum_{\lambda} \mathbf{m}(\lambda) \odot \mathbf{X}(\lambda) + \mathbf{G} \tag{2}$$

Where $\mathbf{Y}\in \mathbb{R}^{H\times W}$ is the two-dimensional data received by the sensor satisfies through a three-dimensional mask $\mathbf{M}\in \mathbb{R}^{H\times W\times N_\lambda}$ and measurement noises $\mathbf{G}\in \mathbb{R}^{H\times W}$, $\mathbf{m}(\lambda)\in \mathbb{R}^{H\times W}$ denotes the equivalent mask at the representative wavelength λ, which is determined by the properties of the filters and their spatial arrangement. Compared to the traditional CASSI system, the optical filter-based HSI system achieves the encoding process of a three-dimensional mask solely through a physical two-dimensional mask, greatly simplifying the imaging optical path. Furthermore, because different filters have significantly varied responses to different spectral bands, the non-correlation between column vectors of the physical imaging model $\mathbf{M}$ is enhanced, posing higher requirements for the decoupling ability of the reconstruction algorithm.

3.2. Spatial-Spectral Cumulative Attention

The hierarchical attention mechanism was previously proposed for document classification [45]. Given that different documents contain sentences of varied lengths and semantics, extracting key features from each text before classifying the entire content can effectively enhance the final classification accuracy and simplify the network structure and parameter size. Notably, in hyperspectral reconstruction,

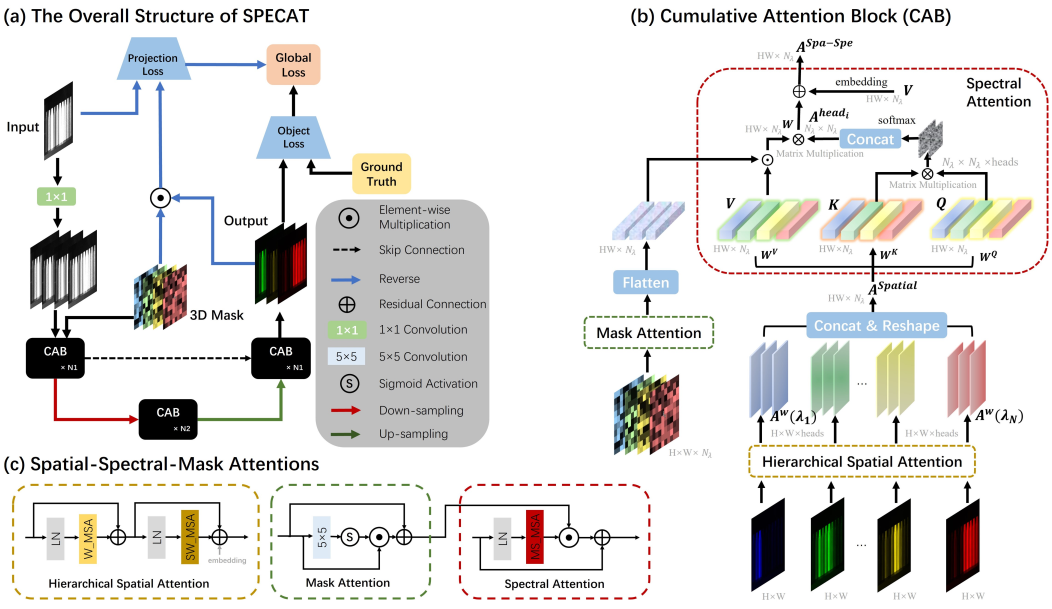


Figure 3. Diagram of the framework structure with SPECAT for HSI reconstruction.

the three-dimensional data is sparse in the spatial domain, indicating the presence of a large amount of redundant information. The reconstruction in the spectral domain, on the other hand, ultimately determines the reliability of the HSI technology. Therefore, inspired by the hierarchical classification method for documents, we propose a Cumulative-Attention Block (CAB) with a hierarchical structure. It first identifies the spatial domain information that best characterizes each spectral band and then uses spectral attention to recover their fine-grained feature details.

As shown in Fig. 3, the CAB consists of a Hierarchical Spatial Attention (HSA) block, a Mask Attention (MA) block and a global spectral attention block. The motivation for introducing spectral attention is inspired by MST [25]. By unfolding the attention in the spectral dimension, it could significantly reduce the overhead of attention computation, and utilize the spectral correlation of hyperspectral data to improve the reconstruction fidelity. The HSA adopts the Swin Transformer[46] attention framework, which currently performs excellently in multi-dimensional image feature extraction. Since the spatial attention between different spectral bands is independent, CAB can focus more on the areas where the spatial domain changes with spectral variations, thus better recovering the spatial domain information during different spectral bands.

The spatial attention considers each local spatial feature map under different spectral bands as a token and calculates self-attention along the spatial dimension. Referring to the SST [27], based on the attention module of the Swin Transformer, we have added the Feature Feed Forward Network (FFN) module and the LayerNorm (LN) layer. The FFN module consists of a series of layers connected in sequence: a 1x1 convolutional layer, a GELU activation layer, a 3x3 convolutional layer, another GELU activation layer, and a final 1x1 convolutional layer. The input $\mathbf{X}$ is computed for Windowed Multi-head Attention (W-MSA) after passing through LN. It is first window-sampled into tokens $\mathbf{x} \in \mathbb{R}^{\frac{H}{s} \cdot \frac{W}{s} \times s \cdot s \times N_\lambda}$, and then linearly projected into queries $\mathbf{Q}(\lambda_i) \in \mathbb{R}^{s^2}$, keys $\mathbf{K}(\lambda_i) \in \mathbb{R}^{s^2}$, and values $\mathbf{V}(\lambda_i) \in \mathbb{R}^{s^2}$ through Eq. (3) with the learnable weights $\mathbf{W}^{\mathbf{Q}(\lambda_i)}(\lambda_i)$, $\mathbf{W}^{\mathbf{K}(\lambda_i)}(\lambda_i)$, and $\mathbf{W}^{\mathbf{V}(\lambda_i)}(\lambda_i) \in \mathbb{R}^{s \times s}$ for different wavelength (λ_i) . Afterward, it goes through LN, and the above process is repeated with shift windows (i.e. SW-MSA). Ultimately, we obtain the window spatial feature output $\mathbf{A}^w(\lambda_i) \in \mathbb{R}^{s^2}$ corresponding to different wavelengths by Eq. (4). Where $\mathbf{B} \in \mathbb{R}^{s^2 \times s^2}$ represents the relative position embedding, and the scaling factor d_k is equal to the dimension of the $\mathbf{K}(\lambda_i)$. Then the $\mathbf{A}^{spatial}(\lambda_i) \in \mathbb{R}^{HW}$ is obtained after concatenation and residual connection following multi-head of the $\mathbf{A}^w(\lambda_i)$.

$$\begin{aligned}\mathbf{Q}(\lambda_i) &= \mathbf{x}(\lambda_i)\mathbf{W}^{\mathbf{Q}(\lambda_i)}(\lambda_i),\\ \mathbf{K}(\lambda_i) &= \mathbf{x}(\lambda_i)\mathbf{W}^{\mathbf{K}(\lambda_i)}(\lambda_i),\\ \mathbf{V}(\lambda_i) &= \mathbf{x}(\lambda_i)\mathbf{W}^{\mathbf{V}(\lambda_i)}(\lambda_i)\end{aligned} \tag{3}$$

$$\mathbf{A}^w(\lambda_i) = \mathbf{V}(\lambda_i)\text{softmax}(\frac{\mathbf{K}(\lambda_i)^T \cdot \mathbf{Q}(\lambda_i)}{\sqrt{d_k}} + \mathbf{B}) \tag{4}$$

$$\mathbf{A}^{spatial}(\lambda_i) = \text{concat}(\mathbf{A}^w(\lambda_i)) + \mathbf{x}(\lambda_i) \tag{5}$$

The spectral attention considers each spectral feature map as a token and computes self-attention along the spectral dimension. Inspired by the MST [25], we first combine

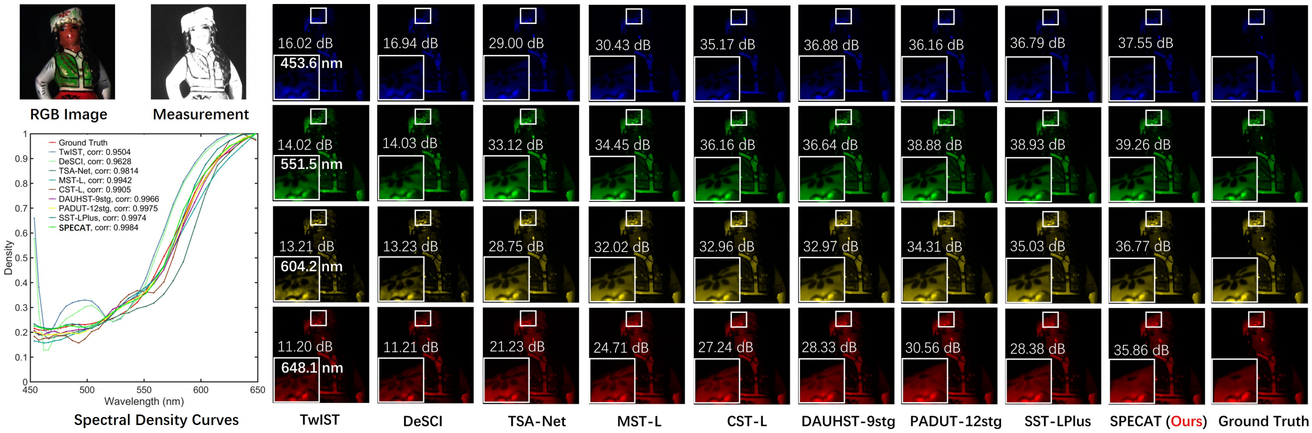


Figure 4. Visual comparisons of our SPECAT and SOTA methods of Scene 6 with 4 out of 28 spectral channels on the KAIST dataset with the optical filter-based HSI system data. The RGB images are generated by three channels [625.0 nm, 551.5 nm, 462.0 nm]. The numbers in the figures represent the PSNR of the enlarged image for details.

and reshape $\mathbf{A}^{spatial}(\lambda_i)$ obtained from different wavelengths λ_i into tokens $\mathbf{A}^{spatial} \in \mathbb{R}^{HW \times N_\lambda}$. Then $\mathbf{A}^{spatial}$ is also linearly projected into queries $\mathbf{Q} \in \mathbb{R}^{HW \times N_\lambda}$, keys $\mathbf{K} \in \mathbb{R}^{HW \times N_\lambda}$, and values $\mathbf{V} \in \mathbb{R}^{HW \times N_\lambda}$through Eq. (6) with the learnable weights $\mathbf{W}^{\mathbf{Q}}$, $\mathbf{W}^{\mathbf{K}}$, and $\mathbf{W}^{\mathbf{V}} \in \mathbb{R}^{N_\lambda \times N_\lambda}$. The $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ are both divided into N heads along the spectral channel dimension.

$$\begin{aligned}\mathbf{Q} &= \mathbf{A}^{Spatial}\mathbf{W}^{\mathbf{Q}},\\ \mathbf{K} &= \mathbf{A}^{Spatial}\mathbf{W}^{\mathbf{K}},\\ \mathbf{V} &= \mathbf{A}^{Spatial}\mathbf{W}^{\mathbf{V}}\end{aligned} \tag{6}$$

To introduce the physical constraints of the mask into the network, inspired by MST [25], we take the product of $\mathbf{Q}$ from the above equation with the MA as the final $\mathbf{Q}$ for computing the space-spectral hierarchical attention. Considering that for the optical filter-based HSI system, the mask itself reflects the characteristics of three-dimensional space-spectral sampling. Therefore, to enhance the network's generalizability across different systems, we employ a simplified mask attention framework. As shown in Fig. 3, its mathematical expression is given by Eq. (7).

$$\mathbf{A}^{mask} = f(\mathbf{M}, \mathbf{W}^{\mathbf{A}^{mask}}) + \mathbf{M} \tag{7}$$

Where f denotes convolutional and activation operations and $\mathbf{W}^{\mathbf{A}^{mask}} \in \mathbb{R}^{N_\lambda \times N_\lambda}$. Finally, the spatial-spectral cumulative-attention feature output for each head $head_i$ and the $\mathbf{A}^{Spa-Spe} \in \mathbb{R}^{HW \times N_\lambda}$ are given by Eq. (8) and Eq. (9). Where σ_i is a learnable parameter to re-weight the multiplication, $\mathbf{W} \in \mathbb{R}^{N_\lambda \times N_\lambda}$ denotes the learnable weight, and P represents the function generating position embeddings, analogous to the method employed in MST++ [44].

$$head_i = \text{softmax}(\sigma_i \mathbf{Q}_i \mathbf{K}_i^T)(\mathbf{V}_i \odot \mathbf{A}^{mask}) \tag{8}$$

$$\mathbf{A}^{Spa-Spe} = \text{concat}_{i=1}^{N}(head_i) \cdot \mathbf{W} + P(\mathbf{V}) \tag{9}$$

3.3. Dual-domain Loss and Projection Constraint

Most existing E2E algorithms primarily rely on designing loss functions for the target domain. Although network training performance can be enhanced through frequency domain transformations and other methods, the lack of spectral compression constraints limits performance. Inspired by [27, 47], to integrate the physical model of spectral reconstruction into the network, enhancing the network's interpretability and reconstruction accuracy while minimizing the additional computational cost, we propose a Dual-domain Loss Function (DLF) for the projection domain and the object domain in this study.

$$\mathcal{L} = ||\mathbf{X}_{out} - \mathbf{X}_0||_2 + ||\sum_{\lambda} \mathbf{M} \odot \mathbf{X}_{out} - \mathbf{Y}_0||_2 \tag{10}$$

Where the $X_0 \in \mathbb{R}^{H \times W \times N_\lambda}$ and $Y_0 \in \mathbb{R}^{H \times W}$ are the true values of the object and projection, respectively. Although the proportion of contributions from these two terms can be adjusted, the ratio chosen in Eq. (10) is based on the consideration that both direct reconstruction of the objection and validation of the imaging process are equally important.

3.4. The Overall Structure of SPECAT

Our proposed SPECAT network's overall framework is illustrated in Fig. 3. To learn features at various scales, we refer to the U-Net structure [48]. The input reconstruction data passes through a series of N_1 CAB blocks, followed by a downsampling (4×4 convolution) that reduces the spatial dimensions to a quarter of the original size. Subsequently, it goes through N_2 CAB blocks, and during upsampling, it is concatenated with shallow features through skip connections before passing through another set of N_1 blocks to output the reconstructed three-dimensional hyperspectral data. We set $N_1 = 2$ and $N_2 = 1$ in the experiments.

4. Experiments

4.1. Experimental Design

In our experiments, we worked with 28 spectral channels ranging from wavelengths of 450 nm to 650 nm. Our experiments were conducted on both simulated and real hyperspectral image datasets.

4.2. Optical Filter-based HSI System Data

We engaged with two distinct datasets: CAVE [51] and KAIST [52]. The CAVE dataset encompasses 32 hyperspectral images, each with a uniform spatial resolution of 512×512 pixels. The KAIST dataset contained 30 hyperspectral images, each with a larger spatial resolution of 2704×3376 pixels. Aligning with the previous methodology [25–29], we designated the CAVE dataset for training purposes. Additionally, we selected 10 scenes from the KAIST dataset for our testing phase. All training and test sets were processed through an optical filter-based mask to generate simulated measurements. In this section, we selected the mask consisting of Fabry-Pérot filters described in [33] for optical filter-based HSI system simulation, due to its representativeness in design and performance.

4.3. Real CASSI System Data

For real HSI data, we employed a dataset acquired using the CASSI system, which was previously developed and described in TSA-Net [26]. All training sets simulated measurements by the CASSI system, with noise added to mimic the projection data acquisition of real sensors.

4.4. Evaluation Metrics

To evaluate the performance of HSI reconstruction, we employed two metrics: peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [53].

4.5. Implementation Specifics

We developed the SPECAT using Pytorch, which was trained on a single Nvidia RTX 4090 GPU. All the models were trained via the Adam[54] optimizer with β_1 set to 0.9 and β_2 to 0.999 for 300 epochs. The initial learning rate was fixed at 4×10^{-4}, with a Cosine Annealing learning rate schedule ($\gamma = 1 \times 10^{-6}$).

5. Results and Discussion

5.1. Results of Optical Filter-based HSI System

A comparison of several state-of-the-art (SOTA) algorithms, including TwIST[21], DeSCI[23], TSA-Net[26], GAP-net[49], BIRNAT[50], MST-L[25], MST++[44], CST-L[28], DAUHST-9stg[29], PADUT-12stg[30] and SST-LPlus[27], was conducted. These algorithms were evaluated under uniform conditions (test size = 256 × 256) for parameters, GFLOPs, PSNR, and SSIM, based on 10 scenes from simulation datasets.

(i) As shown in Tab. 1, our SPECAT algorithm showed superior performance, achieving an average PSNR of 40.37 dB and an average SSIM of 98.6%, outperforming existing models by over 1.24 dB in PSNR and 1.23% in SSIM. It notably surpassed the other algorithms in PSNR and SSIM, demonstrating its high efficiency. A visual comparison of SPECAT with other methods on select spectral channels of Scene 6 in Fig. 4 revealed that SPECAT provided richer spatial details and clearer textures. The spectral curve of SPECAT also indicated higher spectral accuracy and better perceptual quality.

(ii) Quantitatively, SPECAT enhanced performance while reducing memory and computational resource consumption. Compared with other Transformer-based methods like SST-LPlus[27], CST-L[28], and MST-L[25], SPECAT showed higher PSNR with significantly fewer parameters and GFLOPs. It also outperformed DU networks like PADUT-12stg[30] and DAUHST-9stg[29] in PSNR, with much lower parameter and GFLOPs requirements, highlighting its efficiency and effectiveness in terms of PSNR, parameter count, and GFLOPs. As Fig. 1 intuitively illustrates, SPECAT outperforms the aforementioned optimal algorithms [25–30, 41, 42, 48–50, 55, 56] in terms of PSNR, parameter count, and GFLOPs.

5.2. Results of CASSI System

To evaluate the effectiveness of our method on data from CASSI systems, we conducted tests using five compressed measurements captured by an actual CASSI system. All compared methods were trained on the CAVE dataset, adopted the optimal model within allowable parameter and memory ranges, employed the same actual mask, and had 11-bit quantization noise injected for fair comparisons [25]. Fig. 5 visually compares our proposed SPECAT with existing state-of-the-art (SOTA) methods, including TSA-Net[26], GAP-net[49], BIRNAT[50], MST[25], MST++[44], CST[28], DAUHST[29], PADUT[30], and SST-M[27]. When employing the same parameters as that used in the optical filter-based HSI system, SPECAT achieved reconstruction results comparable to SOTA methods. Although there is still room for improvement in the spatial details recovered for the CASSI system, it exhibits significant enhancement in recovering spectral dimension information. In Scene 1, at 503.9nm, the right green flower appears bright while the left red flower nearly disappears, and at 604.2nm, the opposite occurs. The reconstruction results of Our SPECAT best displayed the contrast between the two flowers, demonstrating its superiority in spectral reconstruction accuracy. In Scene 4, SPECAT more clearly reconstructed the pentagram's edge contours under different spectral bands compared to current SOTA methods.

Algorithms	Parameters	GFLOPs	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg
TwIST[21]	-	-	25.10 0.689	23.08 0.620	21.45 0.721	30.16 0.832	21.85 0.652	20.87 0.640	22.45 0.645	21.75 0.653	22.56 0.698	22.73 0.572	23.20 0.678
DeSCI[23]	-	-	27.15 0.753	22.90 0.615	26.78 0.812	34.78 0.893	23.89 0.701	22.42 0.685	24.40 0.681	22.10 0.728	24.50 0.872	23.60 0.590	25.25 0.733
TSA-Net[26]	44.2 M	135.1	32.30 0.936	31.26 0.906	28.53 0.875	36.36 0.931	30.37 0.937	33.06 0.950	31.04 0.891	30.88 0.924	28.99 0.872	32.62 0.947	31.54 0.917
GAP-net[49]	4.26 M	84.5	34.30 0.942	31.59 0.893	28.48 0.831	36.62 0.909	32.74 0.923	34.30 0.921	31.76 0.877	31.52 0.906	30.01 0.867	33.45 0.948	32.48 0.901
BIRNAT[50]	4.35 M	2130	37.58 0.965	38.53 0.964	37.31 0.950	41.39 0.952	37.39 0.969	38.16 0.962	36.61 0.952	35.58 0.950	37.07 0.954	35.93 0.962	37.55 0.958
MST-L[25]	2.03 M	28.5	36.55 0.963	36.29 0.953	33.46 0.904	39.78 0.951	35.40 0.964	36.10 0.963	34.53 0.924	33.08 0.945	34.38 0.926	35.19 0.966	35.48 0.946
MST++[44]	1.34 M	19.6	36.65 0.960	37.14 0.963	34.84 0.936	38.94 0.959	36.44 0.968	36.96 0.969	35.37 0.940	34.27 0.953	34.58 0.941	35.08 0.974	36.03 0.956
CST-L[28]	3.00 M	40.1	36.98 0.963	38.34 0.963	35.89 0.939	40.98 0.951	36.19 0.967	37.23 0.957	35.75 0.944	34.64 0.946	36.41 0.946	35.93 0.961	36.83 0.954
DAUHST-9stg[29]	6.15 M	79.5	38.37 0.972	39.91 0.977	37.71 0.965	42.97 0.967	37.69 0.980	39.05 0.974	37.62 0.964	36.11 0.965	38.45 0.971	37.39 0.976	38.53 0.971
PADUT-12stg[30]	5.38 M	90.5	38.42 0.974	40.34 0.983	38.95 0.972	43.50 0.977	38.22 0.983	39.16 0.979	38.21 0.971	36.03 0.969	39.45 0.979	37.30 0.981	38.96 0.977
SST-LPlus[27]	9.71 M	162.1	39.49 0.977	40.64 0.978	39.92 0.970	42.79 0.976	38.78 0.980	39.34 0.976	38.21 0.968	36.53 0.966	39.47 0.971	36.17 0.970	39.13 0.973
SPECAT	**0.29M**	**12.4**	**40.24** **0.982**	**42.40** **0.986**	**41.43** **0.978**	**44.90** **0.982**	**39.62** **0.987**	**39.90** **0.984**	**39.41** **0.977**	**37.49** **0.977**	**40.45** **0.982**	**37.90** **0.983**	**40.37** **0.986**

Table 1. Comparison of Parameter count, GFLOPs, PSNR and SSIM (upper and lower entry in each cell, respectively) of different methods on 10 simulation scenes (S1∼S10) for optical filter-based HSI system.

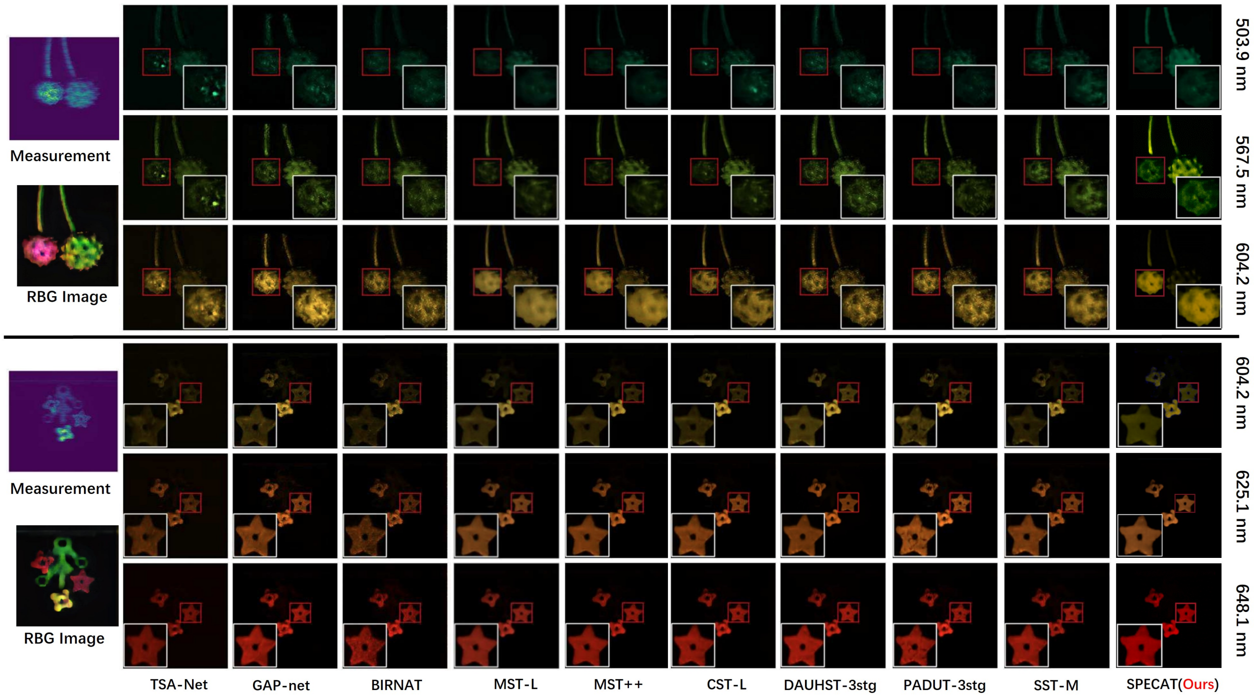


Figure 5. Comparison of real HSI reconstruction for Scene 1 and Scene 4 measured by the CASSI system in [26], with 6 spectra randomly chosen out of 28. The RGB images are generated by three channels [625.0 nm, 551.5 nm, 462.0 nm].

5.3. Ablation Study

As shown in Tab. 2, a break-down ablation study was conducted on SPECAT to investigate the impact of each module on performance enhancement, including the Hierarchical Spatial Attention (HSA) and Mask Attention (MA) in CAB. The setting of the ablation experiment was consistent with the simulation experiment of the hyperspectral system based on optical filters. The baseline is a spectral-wise Transformer with the same architecture with the removal of CAB and DLF. The ablation study demonstrated that the proposed CAB significantly improves the network's reconstruction accuracy. Among them, HSA plays a major role, surpassing networks that only increase the number of spectral attention modules and network depth in terms of improvement (e.g. MST++). Additionally, the results also show that under the same network model and training parameters, after incorporating projection constraints, there was an approximate improvement of 0.4 dB in PSNR on the performance metrics and a slight improvement in SSIM. Thus, the dual-domain loss can effectively utilize the physical model constraints of spectral imaging to further enhance the performance of SPECAT.

Baseline	CAB		DLF	PSNR	SSIM	Params(M)	GFLOPs
	HSA	MA					
✓				34.02	0.931	0.16	5.2
✓	✓			38.87	0.964	0.27	11.4
✓	✓	✓		39.96	0.985	0.29	12.4
✓	✓	✓	✓	40.37	0.986	0.29	12.4

Table 2. Break-down ablation study of SPECAT.

CABs	Depth	AMs	FCLs	PSNR	Params(M)	GFLOPs
[2, 1]	**2**	**15**	**40**	**40.37**	**0.29**	**12.4**
[2, 2]	2	18	48	40.42	0.39	13.9
[2, 2, 1]	3	27	72	40.68	1.17	19.9
[2, 2, 2]	3	30	80	40.70	1.54	21.5
SST-LPlus	3	108	∼432	39.13	9.71	162.1

Table 3. Params and cost analysis of SPECAT.

The results of Tab. 3 indicate that the significant reduction in parameters and costs of SPECAT can be attributed to three aspects: (1) Compared to global self-attention, our hierarchical attention mechanism has lower complexity. (2) The stronger representational capability of the CAB allows for the use of fewer Attention Modules (AMs) and Fully Connected Layers (FCLs). (3) The depth (U-shape stage) is reduced. In the experiments, the parameters and cost of SPECAT significantly increase with the increase of AMs and depth, but the performance improvement is limited. The network structure ([2,1]) chosen in our paper aims to balance performance with the lowest possible cost. The configuration of U-shaped stages and the number of CABs can be tailored to specific tasks to attain peak performance.

5.4. Cumulative Attention

As shown in Fig. 6, during the reconstruction process of SPECAT, CAB can utilize the spatial attention of the lower layer to first identify and 'mark out' the spatial characteristics of the object. This aids the spectral attention of the upper layer in more accurately extracting deep features of the three-dimensional hyperspectral data, especially in spatial areas where the spectral response may undergo significant changes. For the 648.1nm wavelength, the petals on the hat of the model were suppressed through spatial attention. Therefore, in Fig. 4, only the reconstruction results of SPECAT accurately represent the absence of petals in this wavelength range, consistent with the true value.

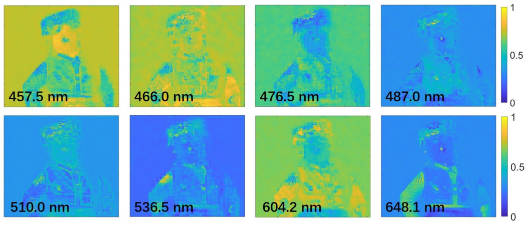


Figure 6. Attention feature maps of different spectral segments obtained through spatial attention during reconstruction. All images have been individually normalized for display purposes.

6. Conclusion

In this study, we propose SPECAT, a spatial-spectral cumulative attention converter specifically designed for high-resolution hyperspectral image reconstruction. The core of SPECAT lies in its CAB, which efficiently extracts detailed features through hierarchical attention and solves the challenge of capturing complex spatial-spectral relationships in hyperspectral images. The experimental results show that SPECAT is superior to SOTA in terms of parameter quantity, computational cost, and spectral reconstruction fidelity. Although there is still room for further improvement in the restoration of spatial details in the reconstruction of CASSI systems, the results of filter-based HSI systems indicate that SPECAT has great potential in advancing on-chip hyperspectral imaging systems and contributing to the miniaturization technology of spectrometers.

Acknowledgments: This work is partially supported by Natural Science Foundation of China (NSFC) under contracts No. 62125106, 62088102 and 62271283, in part by Ministry of Science and Technology of China under contract No. 2021ZD0109901, in part by Young Elite Scientists Sponsorship Program by CAST No. 2021QNRC001.

References

[1] Xueyang Wang, Xiya Zhang, Yinheng Zhu, Yuchen Guo, Xiaoyun Yuan, Liuyu Xiang, Zerun Wang, Guiguang Ding, David Brady, Qionghai Dai, et al. Panda: A gigapixel-level human-centric video dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3268–3278, 2020. 1

[2] Xiaoyun Yuan, Lu Fang, Qionghai Dai, David J Brady, and Yebin Liu. Multiscale gigapixel video: A cross resolution image matching and warping approach. In *ICCP*, pages 1–9. IEEE, 2017.

[3] Jianing Zhang, Tianyi Zhu, Anke Zhang, Xiaoyun Yuan, Zihan Wang, Sebastian Beetschen, Lan Xu, Xing Lin, Qionghai Dai, and Lu Fang. Multiscale-vr: Multiscale gigapixel 3d panoramic videography for virtual reality. In *ICCP*, pages 1–12. IEEE, 2020.

[4] Xiaoyun Yuan, Mengqi Ji, Jiamin Wu, David J Brady, Qionghai Dai, and Lu Fang. A modular hierarchical array camera. *Light: Science & Applications*, 10(1):37, 2021. 1

[5] Tetsuro Ishida, Junichi Kurihara, Fra Angelico Viray, and et al. A novel approach for vegetation classification using uav-based hyperspectral imaging. *Computers and electronics in agriculture*, 144:80–85, 2018. 1

[6] WS Udin, NAS Norazami, N Sulaiman, NA Che Zaudin, S Ma'ail, and AN Mohamad Nor. Uav based multi-spectral imaging system for mapping landslide risk area along jeli-gerik highway, jeli, kelantan. In *2019 IEEE 15th International Colloquium on Signal Processing & Its Applications (CSPA)*, pages 162–167. IEEE, 2019. 1

[7] Stephanie L Wright, Joseph M Levermore, and Frank J Kelly. Raman spectral imaging for the detection of inhalable microplastics in ambient particulate matter samples. *Environmental science & technology*, 53(15):8947–8956, 2019. 1

[8] Simone De Angelis, Eleonora Ammannito, Tatiana Di Iorio, Maria Cristina De Sanctis, Paola Olga Manzari, Fabrizio Liberati, Fabio Tarchi, Michele Dami, Monica Olivieri, Carlo Pompei, et al. The spectral imaging facility: Setup characterization. *Review of Scientific Instruments*, 86(9), 2015. 1

[9] Yunsong Li, Yanzi Shi, Keyan Wang, Bobo Xi, Jiaojiao Li, and Paolo Gamba. Target detection with unconstrained linear mixture model and hierarchical denoising autoencoder in hyperspectral imagery. *IEEE Transactions on Image Processing*, 31:1418–1432, 2022. 1

[10] Min H Kim, Todd Alan Harvey, David S Kittle, Holly Rushmeier, Julie Dorsey, Richard O Prum, and David J Brady. 3d imaging spectroscopy for measuring hyperspectral patterns on solid objects. *ACM Transactions on Graphics (TOG)*, 31 (4):1–11, 2012. 1

[11] Fengchao Xiong, Jun Zhou, and Yuntao Qian. Material based object tracking in hyperspectral videos. *IEEE Transactions on Image Processing*, 29:3719–3733, 2020. 1

[12] Jiayuan Li, Qingwu Hu, and Mingyao Ai. Rift: Multi-modal image matching based on radiation-variation insensitive feature transform. *IEEE Transactions on Image Processing*, 29: 3296–3310, 2019. 1

[13] Shuyang Liu, Hanyou Tian, Chen Zhang, and et al. Towards meat quality determination on miniaturized spectral imaging and computer vision technology. In *Fifth Symposium on Novel Optoelectronic Detection Technology and Application*, volume 11023, pages 381–386. SPIE, 2019. 1

[14] Tingting Liu, Hai Liu, You-Fu Li, Zengzhao Chen, Zhaoli Zhang, and Sannyuya Liu. Flexible ftir spectral imaging enhancement for industrial robot infrared vision sensing. *IEEE Transactions on Industrial Informatics*, 16(1): 544–554, 2019. 1

[15] Minghua Wang, Qiang Wang, and Jocelyn Chanussot. Tensor low-rank constraint and l_0 total variation for hyperspectral image mixed noise removal. *IEEE Journal of Selected Topics in Signal Processing*, 15(3):718–733, 2021. 1

[16] Hao Du, Xin Tong, Xun Cao, and Stephen Lin. A prism-based system for multispectral video acquisition. In *ICCV*, pages 175–182. IEEE, 2009. 1

[17] Ashwin A Wagadarikar, Nikos P Pitsianis, Xiaobai Sun, and David J Brady. Video rate spectral imaging using a coded aperture snapshot spectral imager. *Optics express*, 17(8): 6368–6388, 2009. 1

[18] 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. 1

[19] Lizhi Wang, Zhiwei Xiong, Guangming Shi, Feng Wu, and Wenjun Zeng. Adaptive nonlocal sparse representation for dual-camera compressive hyperspectral imaging. *IEEE transactions on pattern analysis and machine intelligence*, 39(10):2104–2111, 2016. 2

[20] Shipeng Zhang, Lizhi Wang, Ying Fu, Xiaoming Zhong, and Hua Huang. Computational hyperspectral imaging based on dimension-discriminative low-rank tensor recovery. In *ICCV*, pages 10183–10192, 2019.

[21] José M Bioucas-Dias and Mário AT Figueiredo. A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Transactions on Image processing*, 16 (12):2992–3004, 2007. 6, 7

[22] 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.

[23] 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. 6, 7

[24] Ying Fu, Yinqiang Zheng, Imari Sato, and Yoichi Sato. Exploiting spectral-spatial correlation for coded hyperspectral image restoration. In *CVPR*, pages 3727–3736, 2016. 2

[25] Yuanhao Cai, Jing Lin, Xiaowan Hu, and et al. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *CVPR*, pages 17502–17511, 2022. 2, 3, 4, 5, 6, 7

[26] Ziyi Meng, Jiawei Ma, and Xin Yuan. End-to-end low cost compressive spectral imaging with spatial-spectral self-attention. In *ECCV*, pages 187–204. Springer, 2020. 2, 3, 6, 7

[27] Zeyu Cai, Jian Yu, Ziyu Zhang, Chengqian Jin, and Feipeng Da. Sst-reversiblenet: Reversible-prior-based spectral-spatial transformer for efficient hyperspectral image reconstruction. *arXiv preprint arXiv:2305.04054*, 2023. 2, 3, 4, 5, 6, 7

[28] Yuanhao Cai, Jing Lin, Xiaowan Hu, and et al. Coarse-to-fine sparse transformer for hyperspectral image reconstruction. In *ECCV*, pages 686–704. Springer, 2022. 2, 3, 6, 7

[29] 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. 2, 3, 6, 7

[30] Miaoyu Li, Ying Fu, Ji Liu, and Yulun Zhang. Pixel adaptive deep unfolding transformer for hyperspectral image reconstruction. In *ICCV*, pages 12959–12968, 2023. 2, 3, 6, 7

[31] Mikael Marois, Jonathan D Olson, Dennis J Wirth, and et al. A birefringent spectral demultiplexer enables fast hyperspectral imaging of protoporphyrin ix during neurosurgery. *Communications Biology*, 6(1):341, 2023. 2

[32] Jian Xiong, Xusheng Cai, Kaiyu Cui, and et al. Dynamic brain spectrum acquired by a real-time ultraspectral imaging chip with reconfigurable metasurfaces. *Optica*, 9(5):461–468, 2022. 2

[33] Motoki Yako, Yoshikazu Yamaoka, Takayuki Kiyohara, and et al. Video-rate hyperspectral camera based on a cmos-compatible random array of fabry–pérot filters. *Nature Photonics*, 17(3):218–223, 2023. 2, 6

[34] Samuel Ortega, Raúl Guerra, Maria Diaz, and et al. Hyperspectral push-broom microscope development and characterization. *IEEE Access*, 7:122473–122491, 2019. 2

[35] Kuniaki Uto, Haruyuki Seki, Genya Saito, Yukio Kosugi, and Teruhisa Komatsu. Development of a low-cost hyperspectral whiskbroom imager using an optical fiber bundle, a swing mirror, and compact spectrometers. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(9):3909–3925, 2016. 2

[36] Seung-Hwan Baek, Incheol Kim, Diego Gutierrez, and Min H Kim. Compact single-shot hyperspectral imaging using a prism. *ACM Transactions on Graphics (TOG)*, 36(6):1–12, 2017. 2

[37] Kristina Monakhova, Kyrollos Yanny, Neerja Aggarwal, and Laura Waller. Spectral diffusercam: lensless snapshot hyperspectral imaging with a spectral filter array. *Optica*, 7(10):1298–1307, 2020. 2

[38] Zongyin Yang, Tom Albrow-Owen, Weiwei Cai, and Tawfique Hasan. Miniaturization of optical spectrometers. *Science*, 371(6528):eabe0722, 2021. 2

[39] Xuanyu Zhang, Yongbing Zhang, Ruiqin Xiong, Qilin Sun, and Jian Zhang. Herosnet: Hyperspectral explicable reconstruction and optimal sampling deep network for snapshot compressive imaging. In *CVPR*, pages 17532–17541, 2022. 3

[40] Zeyu Cai, Chengqian Jin, and Feipeng Da. Dmdc: Dynamic-mask-based dual camera design for snapshot hyperspectral imaging. *arXiv preprint arXiv:2308.01541*, 2023. 3

[41] Xiaowan Hu, Yuanhao Cai, Jing Lin, and et al. Hdnet: High-resolution dual-domain learning for spectral compressive imaging. In *CVPR*, pages 17542–17551, 2022. 3, 6

[42] Tao Huang, Weisheng Dong, Xin Yuan, and et al. Deep gaussian scale mixture prior for spectral compressive imaging. In *CVPR*, pages 16216–16225, 2021. 3, 6

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, and et al. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[44] Yuanhao Cai, Jing Lin, Zudi Lin, and wt al. Mst++: Multi-stage spectral-wise transformer for efficient spectral reconstruction. In *CVPR*, pages 745–755, 2022. 3, 5, 6, 7

[45] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016. 3

[46] Ze Liu, Yutong Lin, Yue Cao, and et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 4

[47] Hui Xie, Zhuang Zhao, Jing Han, Fengchao Xiong, and Yi Zhang. Dual camera snapshot high-resolution-hyperspectral imaging system with parallel joint optimization via physics-informed learning. *Optics Express*, 31(9):14617–14639, 2023. 5

[48] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 5, 6

[49] Ziyi Meng, Shirin Jalali, and Xin Yuan. Gap-net for snapshot compressive imaging. *arXiv preprint arXiv:2012.08364*, 2020. 6, 7

[50] Ziheng Cheng, Bo Chen, Ruiying Lu, and et al. Recurrent neural networks for snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2264–2281, 2022. 6, 7

[51] Jong-Il Park, Moon-Hyun Lee, Michael D Grossberg, and Shree K Nayar. Multispectral imaging using multiplexed illumination. In *ICCV*, pages 1–8. IEEE, 2007. 6

[52] Inchang Choi, MH Kim, D Gutierrez, DS Jeon, and G Nam. High-quality hyperspectral reconstruction using a spectral prior. Technical report, 2017. 6

[53] Xing Lin, Yebin Liu, Jiamin Wu, and Qionghai Dai. Spatial-spectral encoded compressive hyperspectral imaging. *ACM Transactions on Graphics (TOG)*, 33(6):1–11, 2014. 6

[54] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[55] Xin Miao, Xin Yuan, Yunchen Pu, and Vassilis Athitsos. l-net: Reconstruct hyperspectral images from a snapshot measurement. In *ICCV*, pages 4059–4069, 2019. 6

[56] Lizhi Wang, Chen Sun, Maoqing Zhang, Ying Fu, and Hua Huang. Dnu: Deep non-local unrolling for computational spectral imaging. In *CVPR*, pages 1661–1671, 2020. 6