Aperture Diffraction for Compact Snapshot Spectral Imaging

Tao Lv, Hao Ye, Quan Yuan, Zhan Shi, Yibo Wang, Shuming Wang*, Xun Cao*
Nanjing University, Nanjing, China
{lvtao, yehao, yuanquan, zhanshi, ybwang}@smail.nju.edu.cn
{wangshuming, caoxun}@nju.edu.cn

Abstract

We demonstrate a compact, cost-effective snapshot spectral imaging system named Aperture Diffraction Imaging Spectrometer (ADIS), which consists only of an imaging lens with an ultra-thin orthogonal aperture mask and a mosaic filter sensor, requiring no additional physical footprint compared to common RGB cameras. Then we introduce a new optical design that each point in the object space is multiplexed to discrete encoding locations on the mosaic filter sensor by diffraction-based spatial-spectral projection engineering generated from the orthogonal mask. The orthogonal projection is uniformly accepted to obtain a weakly calibration-dependent data form to enhance modulation robustness. Meanwhile, the Cascade Shift-Shuffle Spectral Transformer (CSST) with strong perception of the diffraction degeneration is designed to solve a sparsity-constrained inverse problem, realizing the volume reconstruction from 2D measurements with Large amount of aliasing. Our system is evaluated by elaborating the imaging optical theory and reconstruction algorithm with demonstrating the experimental imaging under a single exposure. Ultimately, we achieve the sub-super-pixel spatial resolution and high spectral resolution imaging. The code will be available at: https://github.com/Krito-ex/CSST.

1. Introduction

Snapshot spectral imaging (SSI) refers to the acquisition of a 3D spatial-spectral data cube containing spectral information at each spatial location in a single exposure [1]. Whereas spectrum is a fundamental property that covers the physical characteristics of scenes, the visual and discriminative capabilities along the spectral and temporal dimensions will lead to unparalleled high-dimensional visual capability [2]. Hence, the acquisition of high temporal-spatial-spectral resolution data can provide a more comprehensive and refined observation and measurement of dynamic objects or processes.

*Corresponding author

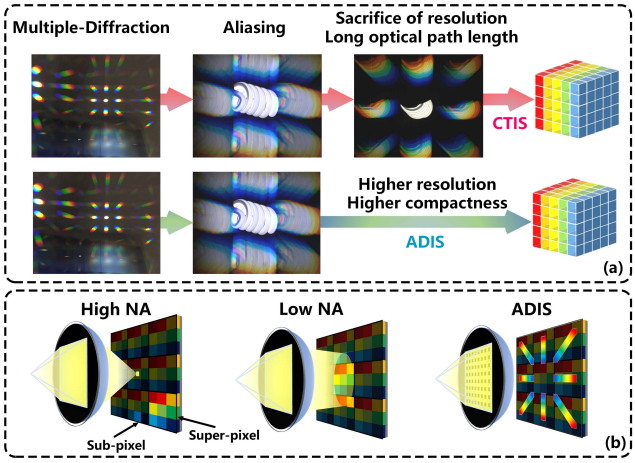

Figure 1. (a) illustrates the CTIS acquisition method and strategy of using long optical path with sacrificing spatial resolution, while ADIS reconstructs from aliasing; (b) depicts different imaging methods for mosaic filter sensors.

Compared with scanning strategies of traditional imaging spectrometers along the spatial or spectral dimension, SSI methods perform specific system designs [3, 4, 5, 6] based on the intrinsic sparsity of the spatial-spectral information of a scene through predefined and well-calibrated modulation or projection paradigms, which can achieve video-level capture of spectral data and have the potential for a wide range of applications in various scenarios such as combustion dynamics [7], cellular dynamics [8], industrial monitoring [9].

However, shortcomings in the compactness, the spatial-temporal-spectral resolution of the imaging system, and the robustness of the modulation limit the application of SSI where portability is paramount [10, 11]:

SSI systems based on computational imaging methods recover the spectral cube by encoding the incident light and solving an underdetermined, sparsity-constrained inverse problem. However, the current prevailing designs rely on bulky relay systems, prisms, or other dispersive elements that result in massive and complex optical systems [10]. Among these, dispersive methods exemplified by CTIS [4]

obviate the need for spatial modulation at the relay system's focal plane, offering the potential for compact design. However, as shown in Figure 1(a), CTIS takes measures of long optical length and sacrifices spatial resolution to reduce the degree of data aliasing. In contrast, we propose a framework that utilizes a single mask at non-focal plane locations to achieve diffractive effects previously accomplished with complex gratings. Specific orthogonal mask diffraction can generate multiplexed spatial spectral projections to reconstruct 3D data cubes without sacrificing system integration, which consists of two sets of parallel lines with orthogonal directions. Overall, ADIS greatly improves the compactness of spectral imaging systems with the same level of integration and manufacturing cost as common RGB or monochrome cameras.

The filter array-based SSI schemes have a compact architecture, but as shown in Figure 1(b), the filter array itself is a sampling trade-off in spatial-spectral dimensions, sacrificing the spatial or spectral resolving ability of imaging systems [12]. The encoding potential of the filter array, however, opens the door to an inverse solution process in ADIS. So a novel encoding scheme is adopted, treating the filter array as a sub-super-resolution encoding array with periodicity. Further, we establish a Transformer-based deep unfolding method, CSST, with orthogonal degradation perception that simultaneously captures local contents and non-local dependencies to meet the challenge of reconstructing convincing sub-super-resolution results from highly confounded measurements.

Additionally, existing SSI technologies rely on multiple optical devices to complete optical encoding in physical space, and its accuracy in practical applications depends on the spatial-spectral mapping relationship determined by the calibration position of optical components, while the ADIS proposed maintains spatial invariance. Under arbitrary perturbation to the aperture mask, it still uniformly maintains the mixed spectral encoding generated by the optical multiplexer to solve the movement problem faced in actual measurement. Furthermore, when the physical parameters of the optical device are determined, the distance between the optical combiner and the sensor is the only variable that affects the spectral mapping. Therefore, ADIS reconstruction only relies on the constant parameters of the system and the distance between the system and the imaging plane, without any complicated calibration.

In summary, specific contributions are:

• A novel SSI framework with an optical multiplexer, enabling high-fidelity and compact snapshot spectral imaging, offering greater resilience against extraneous perturbations.

• A novel diffraction-projection-guided algorithm for hyperspectral reconstruction, capturing the intricate dependencies of diffraction-guided spatial-spectral mapping.

• A prototype device demonstrating excellent hyperspectral acquisition and reconstruction performance.

• Theoretical derivation, structural analysis and necessary trade-offs for system and algorithm design.

2. Related Work

Coded aperture methods involve the utilization of a coded aperture, which selectively blocks incident light either in the focal plane or the rainbow plane [13]. Over the past few decades, various representative systems such as CASSI [3, 14] ,PMVIS [5] and HVIS [15] have been developed to code the light field in the image plane using an occlusion mask, while employing the dispersive element to realize spectral modulation. Additionally, several improvement schemes have been proposed [16, 17, 18], to enhance the effectiveness of the coding process. Despite their efficacy, these systems suffer from the limitations of a bulky optical relay system and the lack of robustness in calibration due to environmental disturbances. In contrast, our system highlights the modulation robustness, which is achieved through the utilization of a clean architecture comprising solely of a mosaic sensor and lenses in combination with an optical multiplexer.

Dispersive methods use prisms or diffractive optics to encode the spectral information. For example, CTIS [4] sacrifices spatial resolution for spectral resolution, which suffers from cone loss; or uses a single dispersion to blur the scene, but leads to a highly ill-conditioned problem and low reconstruction accuracy because the spectral encoding is only at the edges of the objects in the scene [19]; or to further improve the compactness of the system, uses diffractive optics such as DOE to reconstruct 3D hyperspectral information based on the sparsity assumption. However, the modulation robustness of these approaches is still limited by the created anisotropic PSF [20]. in contrast, our system preserves the system compactness with a good enhancement for potential for portable application scenarios.

Filter-array-based methods commonly recover desired channels by utilizing tiled spectral filter arrays in conjunction with the sensor, which incorporate a unique layout of super-pixels periodically arranged in the plane, leading to a reduction in spatial resolution with an increase in the number of sampled channels [12]. While some demosaic techniques may be used in combination with filter-array-based methods, they rely on data that is not initially captured by the sensor [21]. Although constrained by detector and filter dimensions, narrow-band filter-based spectrometers possess a distinct advantage in terms of miniaturization [10]. Various design solutions, such as thin-films [22], planar photonic crystals [23], metasurfaces [24], have been demonstrated in laboratory settings for the development of filter-array-based spectrometers. In this study, we utilize an orthogonal mask to multiplex information from a single point to different sensor locations for encoding purposes. Fur-

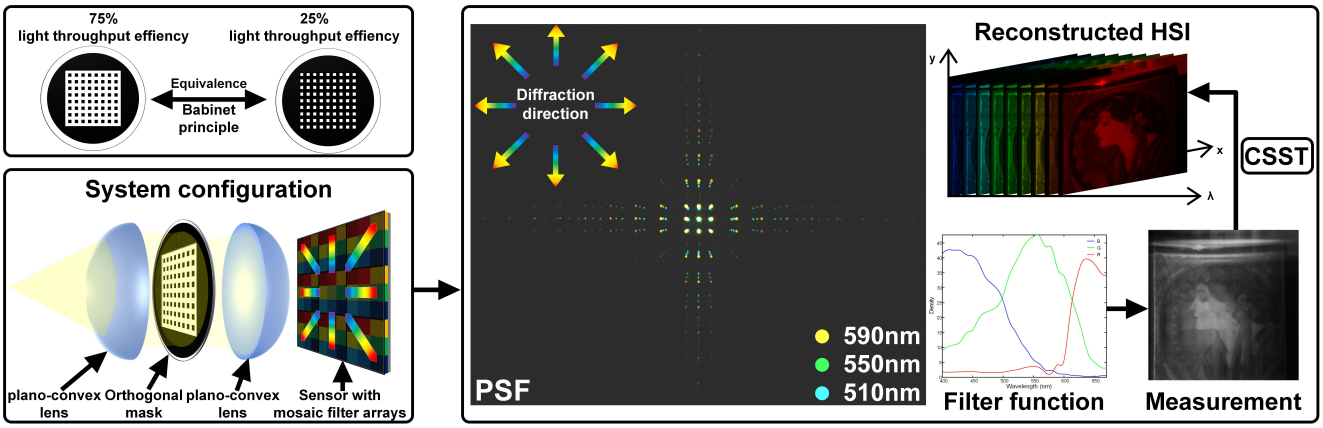

Figure 2. Illustration of ADIS architecture and reconstruction pipeline. In the upper-left, the equivalence between the two complementary masks is depicted. The PSF show in the middle is obtained by ADIS through monochromatic illumination and multiband superimposition.

thermore, our approach can be applied to other hardware solutions for mosaic encoding designs, thus extending its potential applications.

Reconstruction Algorithm. In the field of hyperspectral image (HSI) reconstruction, traditional iterative decoding approaches encounter significant challenges in terms of the time-consuming encoding process and the requirement for prior knowledge [25, 26]. To address these challenges, end-to-end deep-learning approaches have been proposed and have demonstrated remarkable potential in optimizing complex ill-posed problems in various snapshot imaging systems [27, 28, 29, 30]. Notably, λ-net [29] and TSA-net [28] have proposed dual-stage generative models and self-attention, respectively, to map HSI images from a single-shot measurement while modeling spatial and spectral correlation with reasonable computation cost. Recently, Transformer-based methods [31, 32, 33] have emerged as superior alternatives, outperforming previous methods and greatly improving reconstruction efficiency. Additionally, some studies have combined the strengths of both iterative and deep-learning methods by utilizing deep unfolding networks for HSI reconstruction [34, 33]. However, most of these methods rely on a structural mathematical representation of the inverse process, which is absence in ADIS, making the above methods inapplicable or ineffective, so a Transformer-based deep unfolding method, CSST, is designed to cater the requirements of ADIS inverse solving.

3. System overview

This section introduces the proposed SSI system, ADIS, covering its basic configuration, principles, and mathematical logic for determining the system imaging model and device parameters. We also discuss design trade-offs of system parameters and analyze the system's robustness to external perturbations.

3.1. System Configuration

Figure 2 illustrates the configuration of our aperture diffraction imaging spectrometer system, comprising a special lens featuring an orthogonal mask on the principal plane. Alternatively, the lens can be substituted with two plano-convex lenses and orthogonal masks during experimentation. The system is completed with a mosaic array filter camera. When a field point with a smooth reflectance distribution is captured, the system disperses spectral information across different spectral bands in an orthogonal pattern. This pattern directs the information to various sub-pixel positions on the mosaic filter-encoded array. As a result, each sub-pixel on the sensor collects different bands from different spatial positions, enabling sub-super pixel resolution snapshot spectral imaging.

3.2. Imaging Forward Model

We now consider a multi-slit diaphragm has N parallel rectangular diaphragms with rectangular square apertures of width a and length b. The distance between two adjacent slits is d. A simplified schematic of ADIS is shown in Figure 3(a). According to the Huygens-Fresnel principle, each point on a wavefront can be considered a new secondary wave source. Thus, we can treat each rectangular square aperture as a point source for a multi-slit diaphragm. Through the amalgamation of waves generated by each of these point sources, we can effectively derive the complete wave pattern of the entire diaphragm:

$$E_p = E_0 \frac{\sin\beta_1}{\beta_1} \frac{\sin N\gamma_1}{\sin\gamma_1} \frac{\sin\beta_2}{\beta_2} \frac{\sin N\gamma_2}{\sin\gamma_2} \tag{1}$$

Where θ_1 and θ_2 are the diffraction angles in x- and y-directions respectively, $\beta_1 = \frac{1}{2}kb\sin\theta_1$, $\beta_2 = \frac{1}{2}ka\sin\theta_2$, $\gamma_1 = \frac{1}{2}kd\sin\theta_1$, $\gamma_2 = \frac{1}{2}kd\sin\theta_2$. Further, by utilizing the paraxial approximation in far-field imaging, the angular relationship can be transformed into a position relationship ($\sin\theta_1 \approx \tan\theta_1 = \frac{x}{f_2}$, $\sin\theta_2 \approx \tan\theta_2 = \frac{y}{f_2}$). As a result,

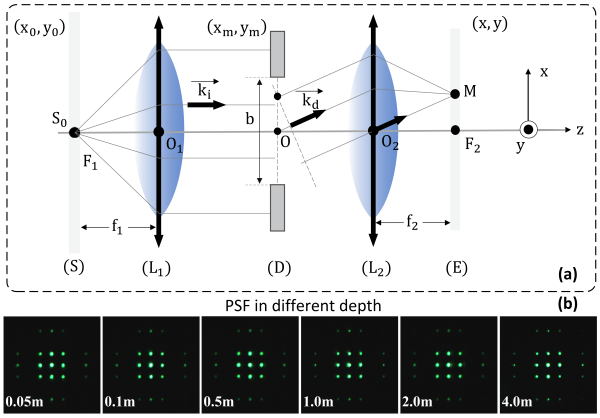

Figure 3. (a) illustrates the simplified schematic of the ADIS's profile; (b) shows the PSF of the system at different depths.

the intensity and position relationship of the diffraction pattern can be represented as follows:

$$I(x, y, \lambda) = I_0 \cdot D(x, y, \lambda) \cdot P(x, y, \lambda) \tag{2}$$

$$D(x, y, \lambda) = \sin c^2(\frac{b}{\lambda f_2} x) \sin c^2(\frac{a}{\lambda f_2} y) \tag{3}$$

$$P(x, y, \lambda) = \left[\frac{\sin(N\frac{\pi d}{\lambda f_2} x)}{\sin(\frac{\pi d}{\lambda f_2} x)}\right]^2 \left[\frac{\sin(N\frac{\pi d}{\lambda f_2} y)}{\sin(\frac{\pi d}{\lambda f_2} y)}\right]^2 \tag{4}$$

Where the formula $D(x, y, \lambda)$ is the diffraction factor describes the diffraction effect of each rectangular square hole. $P(x, y, \lambda)$ is the interference factor describes the effect of multi-slit interference. (x, y) denotes the spatial coordinates on the receiving screen, while f_2 denotes the distance between the diffraction array and the sensor.

Therefore, Given our design with a lens generating orthogonal diffraction in front of the sensor, the forward model of ADIS can be considered as a combination of projection modulation and intensity encoding:

$$L[x, y] = \sum_{\lambda=0}^{K-1} F_\lambda[x, y] \cdot Q[x, y, \lambda] \tag{5}$$

Where $F_\lambda[x, y]$ denotes the modulation of optical multiplexer, which is comprehensively conveyed via Equation 2, while $Q[x, y, \lambda]$ denotes filtering and coding influence of mosaic filter sensors.

3.3. Orthogonal Mask Parameters

Through the analytical formula of aperture diffraction in the image plane, we can analyze the relationship between different diffraction orders and the parameters of the mask. We adjust the aperture mask parameters to increase the intensity of first-order diffraction while suppressing low-order diffraction. Increasing the diffraction intensity of one order can add more spectral information to the image plane, while suppressing other diffraction orders can reduce the stray intensity information during image processing. Whereas the dispersion function of the aperture mask is uniquely determined by the imaging focal length f_2 and period d, where the dispersion distance for the first-order diffraction is: $\Delta x_m = \frac{f_2}{d}(\lambda_{\max} - \lambda_{\min})$. Notably, expanding dispersion distance enhances system spectral resolution, yet escalating it also magnifies PSF dispersion, exacerbating reconstruction underdetermination. Combining the effects of dispersion distance and PSF discretization on the reconstruction of the system, we choose appropriate square holes period $d = 10\mu m$, which is within our manufacturing capability. We calculate and compare the intensity distributions of zero-order and first-order diffraction.

For the zero-order diffraction:

$$I = I_0\left(\frac{\sin\beta_1}{\beta_1}\right)^2\left(\frac{\sin\beta_2}{\beta_2}\right)^2 N^4 = I_0 N^4 \tag{6}$$

For the first-order diffraction:

$$I' = I_0 N^4\left(\frac{\sin\beta_1}{\beta_1}\right)^2 = I_0 N^4\left[\frac{d}{b\pi}\sin\left(\frac{b}{d}\pi\right)\right]^2 \tag{7}$$

Let $m = \frac{d}{b}$, So $I'/I_0 = \left[\frac{m}{\pi}\sin\left(\frac{\pi}{m}\right)\right]^2$. According to calculations, the intensity contrast between the zero-order diffraction and the first-order diffraction depends entirely on $\frac{d}{b}$, which is the ratio between the opening aperture and the spacing of the square holes.

Furthermore, considering the diffraction pattern defined in Equation 2, varying m also influences diffraction patterns. And when we determine $a = b = 5\mu m$ for the case of $d = 10\mu m$, all the even orders will missing, which can be to reduce the projection complexity appropriately. $I_{D_x - D_y} = I_0 N^4 A_{D_X} A_{D_y}$ is expressed as the intensity relation of different diffraction levels, D_x, D_y denote the number of diffraction orders in the orthogonal direction respectively.

$$A_{D_x} = \begin{cases} 1 & , D_x = 0 \\ \dfrac{4}{D_x^2 \pi^2} & , D_x = 1, 3, 5, ... \\ 0 & , D_x = 2, 4, 6, ... \end{cases} \tag{8}$$

A_{D_x}, A_{D_y} have the same mathematical form and together define the projection form of ADIS. Furthermore, the complementary form of the $N \times N$ square aperture array can be employed using the Babinet principle, thereby elevating the light throughput efficiency from 25% to 75%.

3.4. Modulation Robustness

The maintenance of spatial invariance in optical systems is an indispensable characteristic for effectively addressing interference-related issues. While depth invariance is the main part to be considered in an imaging system, here we address the depth invariance of ADIS. Suppose a monochromatic incident wave field u_0 with amplitude A_0, phase ϕ_0 passes through the optical multiplexer:

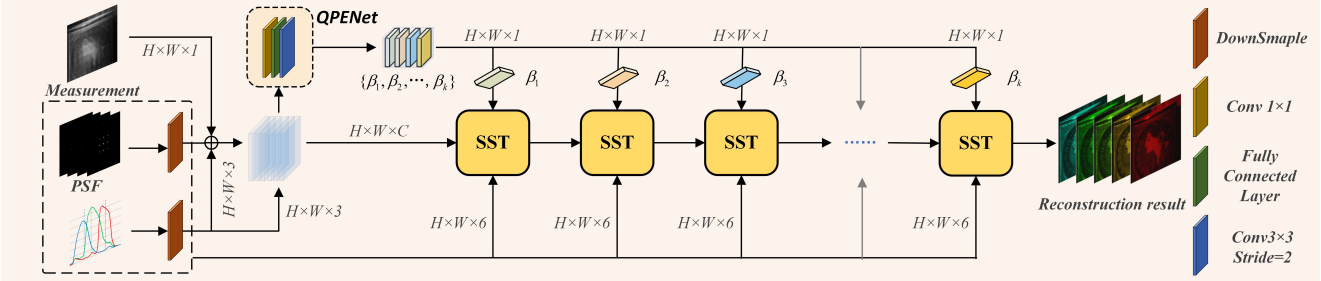

Figure 4. Illustration of COPF architecture with k stages. Theoretically, the SST in the COPF can be replaced with a different denoiser.

$$u_0(x_m, y_m) = A_0(x_m, y_m)e^{i\phi_0(x_m, y_m)} \tag{9}$$

An amplitude encoding and phase shift occurs by the optical multiplexer:

$$u_1(x_m, y_m) = u_0(x_m, y_m)A_1(x_m, y_m)e^{i\phi_1(x_m, y_m)} \tag{10}$$

And when the ADIS is illuminated by a point light source located at depth Z. The spherical wave filed u_0 emitted by the source incident to the optical multiplexer can be represented by:

$$u_0(x_m, y_m; Z) \propto \frac{1}{\xi}e^{ik(\xi - Z)} \tag{11}$$

Where $\xi = \sqrt{{x_m}^2 + {y_m}^2 + Z^2}$.Since the aperture size is negligibly smaller than the imaging depth, the following relationship exists: $\xi \approx Z$. Then the wave field u_1 modulated by the optical multiplexer can be expressed as:

$$u_1(x_m, y_m; Z) \propto \frac{1}{Z}A_1(x_m, y_m)e^{i\{k(\xi - Z) + \phi_1(x_m, y_m)\}} \tag{12}$$

Since $\xi \approx Z$, the point source is relatively close to optical infinity, and $(\xi - Z) \ll \phi_1(x_m, y_m)$ holds in Equation 12. Then Equation 12 can be approximated as Equation 10.The above derivation verifies the depth invariance of the ADIS in a specific depth range. This derivation confirms ADIS's depth invariance within a specific range. We validated this by capturing ADIS PSFs at various depths using a $550nm$ laser (Figure 3(b)), revealing consistent invariance beyond the imaging focal length.

Moreover, ADIS demonstrates resilience to (x, y)-direction device perturbations, provided the modulation plane remains within the imaging optical path. Here, we assume a positional shift p in the y-direction for the Mask. Then we can get: $E_p = E_0 \frac{\sin\beta_1}{\beta_1}\frac{\sin N\gamma_1}{\gamma_1}\frac{\sin\beta_2}{\beta_2}\frac{\sin N\gamma_2}{\gamma_2} \cdot e^{ikp\sin\theta}$. Taking the amplitude of the electric field can obtain $|E_p| = E_0 \frac{\sin\beta_1}{\beta_1}\frac{\sin N\gamma_1}{\gamma_1}\frac{\sin\beta_2}{\beta_2}\frac{\sin N\gamma_2}{\gamma_2}$, which verifies the modulation robustness of the system.

4. Hyperspectral Reconstruction

Drawing on the benefits of self-attention for simultaneously capturing both short- and long-ranged dependencies and dynamic weighting, the Transformer architecture has demonstrated exceptional performance in a range of tasks [31, 32, 33, 35, 36, 37]. In parallel, the deep unfolding framework shows considerable promise through the utilization of multi-stage networks to map measurement outcomes onto the HSI, coupled with layer-by-layer optimization of the imaging system's priori model. This approach affords a more seamless integration between the depth unfolding framework and the imaging model.

In this paper, we present the Cascade Shift-Shuffle Spectral Transformer (CSST) algorithm, which is designed to improve network degradation perception by leveraging shift and shuffle operations that conform to the physical model of imaging and possess a strong perception of orthogonal diffraction projection patterns.

4.1. COPF

To tackle the aforementioned challenges, we develop a Cascaded Orthogonal Perception Framework (COPF) that utilizes a deep unfolding framework to address the aperture diffraction degradation process. The COPF is illustrated in Figure 4. First, a lightweight Quantitative Parameter Estimation network (QPENet), is designed to estimate key cues for later iterations from the system's measurements and priori information such as filter-encoded spectral response and orthogonal diffraction patterns. Notably, the computed PSF exhibits greater spatial extent than the input filter function. To tackle data redundancy, we first downsample the PSF's spatial resolution and the filter function's channel dimension. Figure 4 illustrates the architecture of QEPNet, which includes a $conv1\times1$, a $conv3\times3$, and three fully connected layers. The estimated parameter $\beta = \{\beta_1, \beta_2, ..., \beta_k\}$ is a multichannel feature map that has the same resolution as the input features, whose number of channel layers is kept consistent with the number of iterations, allowing the estimated parameters to guide and optimize the reconstruction process pixel by pixel. Subsequently, COPF adaptively adjusts the feature map information to guide the iterative learning by inputting β channel by channel at different levels of iterations. The initial values for the iterative process in COPF are acquired through a multi-scale integration of system measurements and prior knowledge. During the iterative learning process, the denoiser is cascaded with different cue information input directly in the iterative framework to fully utilize the guiding role of β.

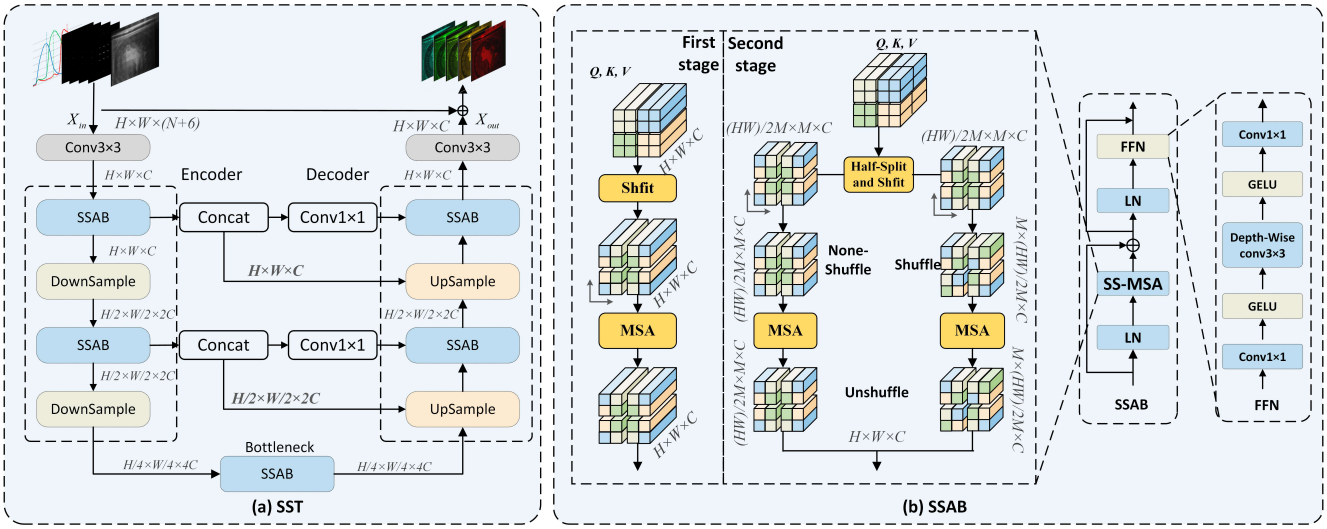

Figure 5. (a) Diagram of SST with three-layer U-shaped structure; (b) SSAB consists of a FFN, a SS-SMA and two LN layers.

4.2. Shift-Shuffle Transformer

The utilization of transformer models for global and local perception encounters challenges of a restricted receptive field and computationally intensive processing. So we propose a novel denoiser, Shift-Shuffle Transformer (SST) as shown in Figure 5, to be inserted in COPF. SST employs channel-shuffle and shift operations with fixed length (set to 1), introduced at the feature map level in the orthogonal direction. These operations improve the model's ability to perceive blending generated by aperture diffraction, while also facilitating the modeling of both short and long distances via the shift operation's function as a special token mixer. It is worth noting that the incorporation of the shift operation does not result in an increase in the total number of algorithm parameters.

Similar to [31, 33], we utilize a three-layer U-shaped structure as the base framework of SST as shown in Figure 5(a). Firstly, SST uses a $conv3 \times 3$ to map reshaped input X_k concatenated with stretched β_k, filter function $\sigma \in \mathbb{R}^{H\times W\times 3}$, PSF $\varsigma \in \mathbb{R}^{H\times W\times 3}$ into feature $X_0 \in \mathbb{R}^{H\times W\times C}$. Secondly, X_0 passes through the encoder, bottleneck, and decoder to be embedded into deep feature$X_f \in \mathbb{R}^{H\times W\times C}$. Basic unit Shift-Shuffle Attention Block (SSAB) assumes the composition of encoder and decoder.

Shift-Shuffle Attention Block consists of two layer normalization (LN), a SS-MSA, and a Feed-Forward Network (FFN) follows the classic design. The most important part of SSAB is Shift-Shuffle Multi-head Self-Attention(SS-MSA) with two stages:

First Stage. In the first stage of SST, only shift operations $\Upsilon(\cdot)$ are performed on the channels. for input tokens $X_{in} \in \mathbb{R}^{H\times W\times C}$:

$$A_1^i = \text{softmax}(\Upsilon(\frac{Q_1^i K_1^{i^T}}{\sqrt{d_h}} + P_1^i))V_1^i \tag{13}$$

Where $h = 1$, $d_h = C$, $\Upsilon(\cdot)$ denotes shifting the input feature map by one pixel in each of its last two dimensions. And the output of first stage is $S(X_{in})_1 = \sum_{i=1}^{h} A_1^i W_1^i$.

Second Stage. Q, K, V will be split into two equal parts along the channel dimension as: $Q_2 = [Q_{2f}, Q_{2s}], K_2 = [K_{2f}, K_{2s}], V_2 = [V_{2f}, V_{2s}]$. The two parts perform different operations separately and get the corresponding results:

$$A_{2f}^i = \text{softmax}(\Upsilon(\frac{Q_{2f}^i K_{2f}^{i^T}}{\sqrt{d_h}} + P_{2f}^i))V_{2f}^i \tag{14}$$

$$A_{2s}^i = \Theta^T(\text{softmax}(\Upsilon(\frac{\Theta(Q_{2s}^i)\Theta(K_{2s}^{i^T})}{\sqrt{d_h}} + P_{2s}^i))\Theta(V_{2s}^i)) \tag{15}$$

Where $h = 1$, $d_h = \frac{C}{2}$, $\Theta(\cdot)$ denotes the channel shuffle operations like ShuffleNet [38] and DAHST [33]. And the output of second stage is:

$$S(X_{in})_2 = \sum_{i=1}^{h} A_{2f}^i W_{2f}^i + \sum_{i=1}^{h} A_{2s}^i W_{2s}^i \tag{16}$$

Then we reshape the result of Equation 16 to obtain the output $X_{out} \in \mathbb{R}^{H\times W\times C}$. The global employment of shift operations, without any supplementary computational overhead, conforms with the ADIS imaging paradigm, while combined with Shuffle operations, enhances the CSST's perceptual capabilities.

5. Experimental analysis

Similar to [28, 42, 43, 44, 31, 32, 33], 28 wavelengths are selected from 450nm to 650nm and derived by spectral interpolation manipulation for the HSI data. However, ADIS creates a wide-area, band-by-band form of PSF, which means that we need HSI of larger spatial size to cre-

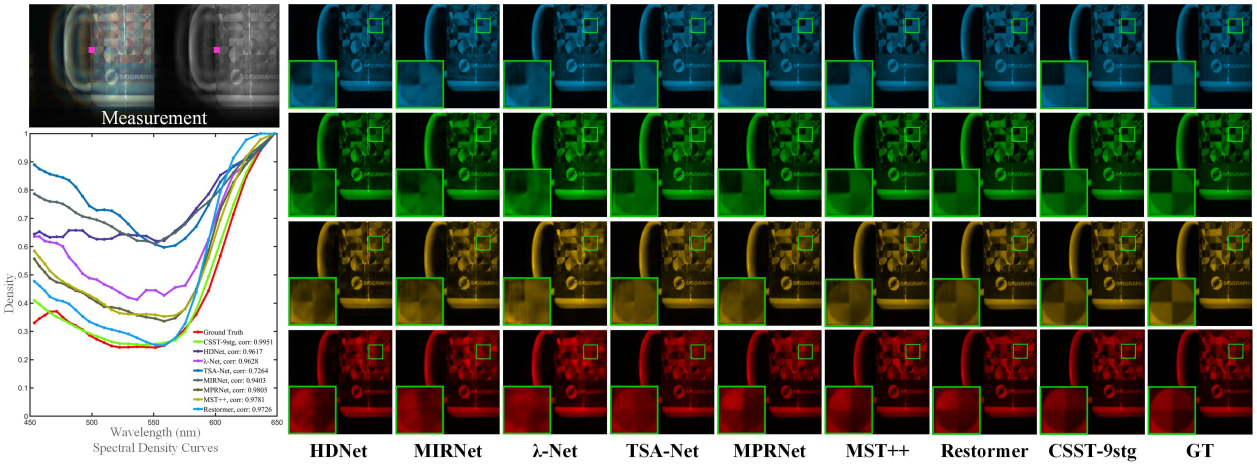

Figure 6. Qualitative comparison of reconstruction results of different algorithms. Zoomed-in patches of the HSI in the fuchsia box are presented in the lower-left of the figure.

Algorithm	Inference Time	Params	GFLOPS	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg
HDNet [30]	2ms	2.37M	144.16	29.55	27.82	24.45	31.38	27.54	27.75	24.43	31.81	33.07	24.13	28.19
				0.879	0.862	0.821	0.883	0.825	0.856	0.812	0.906	0.893	0.834	0.857
MIRNet [39]	2ms	2.04M	14.26	30.266	29.09	25.10	33.04	27.52	28.46	24.66	31.94	33.31	26.22	28.96
				0.907	0.888	0.846	0.909	0.860	0.871	0.822	0.913	0.903	0.854	0.877
lambda-Net [29]	2ms	32.72M	23.10	30.77	28.79	26.73	31.85	28.25	28.69	27.89	32.54	34.76	25.96	29.62
				0.919	0.872	0.851	0.792	0.835	0.827	0.832	0.901	0.909	0.862	0.860
TSA-Net [28]	5ms	44.23M	91.19	32.81	30.26	27.13	34.47	28.58	30.35	26.95	33.98	35.73	26.80	30.71
				0.948	0.923	0.900	0.923	0.901	0.913	0.865	0.941	0.926	0.914	0.915
MPRNet [40]	3ms	2.95M	77.30	32.38	30.91	27.34	34.53	29.24	30.49	28.98	33.97	35.90	27.02	31.08
				0.941	0.931	0.912	0.930	0.907	0.924	0.879	0.942	0.941	0.923	0.923
MST++ [41]	3ms	1.33M	17.45	33.75	31.78	28.87	35.51	29.95	32.34	28.01	35.03	38.53	28.49	32.23
				0.962	0.952	0.942	0.941	0.921	0.948	0.900	0.958	0.960	0.942	0.942
Restormer [35]	10ms	15.12M	87.87	**35.42**	**32.62**	**29.97**	**36.82**	**30.19**	**33.41**	**30.71**	**36.00**	**38.75**	**28.99**	**33.29**
				0.970	**0.959**	**0.951**	**0.942**	**0.926**	**0.956**	**0.909**	**0.961**	**0.962**	**0.945**	**0.948**
CSST-9stg (Ours)	34ms	6.56M	70.44	**34.72**	**34.75**	**31.28**	**36.91**	**31.601**	**33.878**	**30.58**	**36.68**	**39.29**	**31.06**	**34.08**
				0.971	**0.974**	**0.964**	**0.948**	**0.936**	**0.964**	**0.921**	**0.970**	**0.969**	**0.961**	**0.958**

Table 1. Comparison of reconstruction results of different algorithms,Inference time, Params, FLOPS, PSNR (dB) and SSIM are reported.

ate measurements with a certain scale of simulation to conduct experiments. Real experiments and simulation experiments with different methods and different mosaic patterns are conducted.

5.1. simulation Experiments

Simulation Dataset. We adopt two datasets, i.e., CAVE-1024 [28] and KAIST [45] for simulation experiments. The CAVE-1024 consists of 205 HSIs with spatial size 1024×1024 obtained by interpolating and splicing from the CAVE [46] dataset. The KAIST dataset contains 30 HSIs of spatial size 2704×3376. 10 scenes from the KAIST dataset are selected for testing, while the CAVE-1024 dataset and another 20 scenes from the KAIST dataset are selected for training.

Implementation Details. The dispersion step of the primary diffraction is 0.5 spatial pixels, while the simulation experiment is deployed in the range of $400nm$ to $670nm$, which means that $586 \times 586 \times 28$ data cubes are needed to generate 256×256 resolution measurements for conducting experiments while preserving the tertiary diffraction. We implement CSST by Pytorch. All CSST models are trained with Adam [47] optimizer ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) using Cosine Annealing scheme [48] for 300 epochs on an RTX 3090 GPU. The initial learning rate is 4×10^{-4}.

Quantitative Analysis. Table 1 compares the results of CSST and 7 methods including four reconstruction methods(lambda-Net [29], TSA-Net [28], HDNet [30] and MST++ [41]), three Super-resolution algorithms (Restormer [35], MPRNet [40], MIRNet[39]) on 10 simulation scenes. CSST shows the best experimental results on the ADIS spectral reconstruction task, i.e., 34.08dB in PSNR and 0.958 in SSIM. CSST-9stg significantly outperforms two recent SOTA methods Restormer and MST++ by 0.79dB and 1.85dB, demonstrating the effectiveness and acceptability of the imaging system.

Qualitative Analysis. Figure 6 illustrates the comparative performance of our CSST and other methods in the HSI reconstruction of ADIS on the same scene. Visual inspection of the image reveals that the CSST-9stg method provides more intricate details, sharper textures, and well-defined structures. Conversely, the previous approaches produce either overly smooth results that compromise the

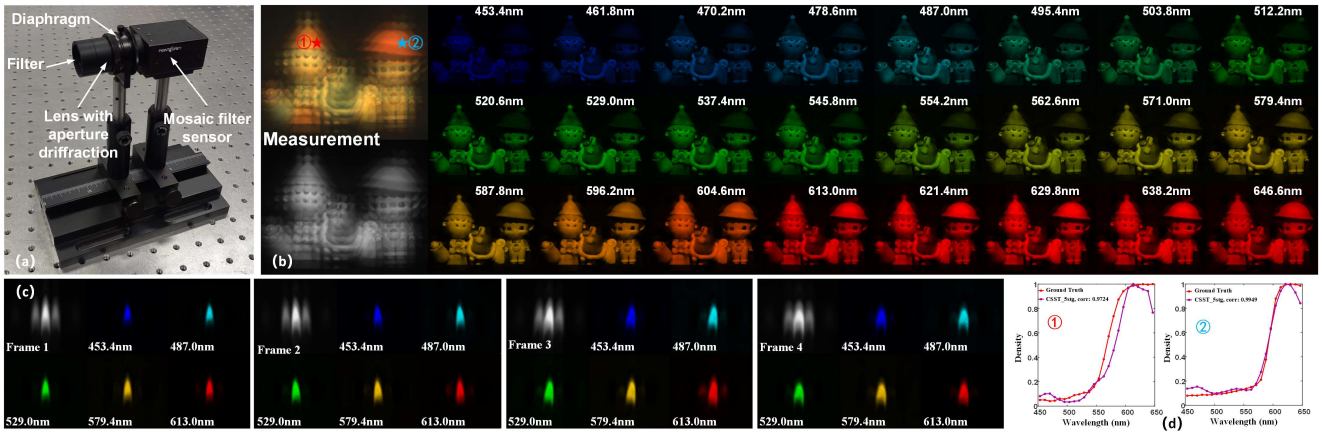

Figure 7. (a) shows the prototype of ADIS; (b) illustrates the ADIS's measurements acquired from real-word and images of different spectral bands recovered by CSST-5stg; (c) shows the measurements and reconstruction results of four frames of a dynamic flame captured by ADIS; (d) Compares the recovered spectral curves and ground truth at the two markers.

underlying structure or introduce color artifacts and speckled textures. Moreover, the lower left corner of the figure presents the spectral profile of the intensity-wavelength corresponding to the fuchsia square. The CSST-9stg spectral profile exhibits the highest correlation and overlap with the reference curve, demonstrating the superiority of our approach in achieving spectral dimensional consistency reconstruction and the effectiveness of ADIS.

5.2. Real HSI Reconstruction

Implementation Details. Firstly, we develop a prototype system utilizing an orthogonal mask with 25% light-throughput and a Bayer array, as illustrated in Figure 6 (top left). This prototype includes additional filters with a wavelength range of $450nm - 650nm$ to restrict the operating band, and an adjustable diaphragm. The small footprint of the system enables high-dimensional information acquisition. The orthogonal mask utilized in prototype is created by overlapping two sets of parallel lines, each with a width and interval of 5µm, and the width uniformity accuracy is 0.2µm. The mask has daiameter of $25.4mm$ and includes a $12mm \times 12mm$ modulation surface. It is custom-priced at \$80 per unit, with costs below \$5 per unit for commercial volume production. Once the physical setup of the system was determined, all projection relationships by could be easily computed by Equation 2 even under the disturbances.

Training Dataset. We trian CSST-5stg with the real configuration on CAVE-1024 and KAIST datasets jointly. Meanwhile, to address the disparity between real-world experiments and simulations arising from inherent noise and our omission of higher-order low-intensity diffraction, we incorporated randomized noise into the training data for model training, thereby bridging the aforementioned gap.

Experimental analysis. The performance of real HSI reconstruction is demonstrated in Figure 7(b), which presents the measurements of a spatial size of 1056×1536 captured from real-world scenes, and the corresponding recovered spectral data of a spatial size of $1056 \times 1536 \times 24$. The reconstructed spectral data exhibit well-structured content, clear textures, and minimal artifacts. Notably, the predicted spectral curves of the two marker points closely match the curves collected using a point spectrometer. These results provide compelling evidence for the correctness and effectiveness of the mathematical model, design framework, and reconstruction algorithm architecture.

Dynamic Performance Verification. In Figure 7(c), the snapshot performance of ADIS is demonstrated through dynamic flame video reconstruction(35 fps).

5.3. Ablation Study

Here we further conduct ablation experiments on each effective component of the CSST algorithm proposed in this paper to demonstrate the necessity of the components used in the algorithm.

COPF	Shift (x,y)	PSNR(dB)	SSIM
✓	✗	27.77	0.870
✓	(1,1)	**28.74**	**0.885**
✗	(1,1)	27.83	0.871
✓	(2,2)	28.44	0.879
✓	(3,3)	28.19	0.873
✓	(4,4)	28.30	0.876
✓	(5,5)	28.01	0.868

Table 2. Break-down ablation results in SS-MSA and COPF, Performance comparison of CSST with different shift step.

We first remove the global shift operations in SS-MSA and COPF from CSST-3stg to conduct the break-down ablation as shown in Table 2. Then We further conduct a comparative analysis to investigate the impact of the shift step size utilized in the shift operations on the effectiveness of CSST reconstruction. The results presented in Table 2 demonstrate a decreasing trend in the reconstruction efficacy of CSST with increasing shift step size. However, it

is noteworthy that all the CSST algorithms with the shift operation outperform the algorithm that lacks the shift operations.

5.4. Simulation with Different Mosaic patterns

The current section aims to investigate the adaptability of the ADIS architecture with diverse mosaic arrays, and here we utilize CSST-5stg for comparative experiments. The experimental setups employed in this study remain consistent with Section 5.1, with the exception of the encoding form of the sensor mosaic array, which is altered. Three distinct mosaics, including a 2×2 pattern with 3 channels, a 3×3 pattern with 4 channels, and a 4×4 pattern with 9 channels, are utilized for comparative experimentation, as demonstrated in Figure 8. With the improvement of filter encoding capability, the imaging performance of ADIS can be further improved.

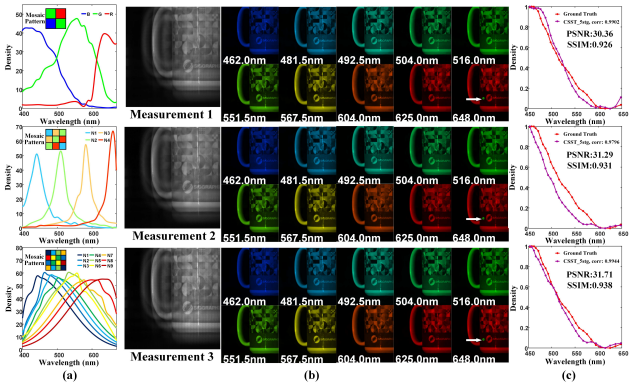

Figure 8. (a) Different mosaic patterns with different filter functions; (b) illustrates the reconstruction results of ADIS combined with different mosaic filter sensor simulations; (c) illustrates recovered spectral curves and ground-truth in the green box.

5.5. Evaluation of Real System

Spectral accuracy. We captured a spectral-interesting, texture-rich scene containing ColorCheck under D65 source illumination to evaluate the spectral accuracy of the hyperspectral image. The measurements captured by our prototype camera and the reconstruction result in the $495.4nm$ channel is shown in Figure 9(a). We also demonstrate excellent agreement between the reconstructed spectra at heart-shaped markers with intricate texture details and the corresponding ground truth spectra.

Spatial resolution. In Figure 9(b), measurements of letters on the ColorCheck are compared with the reconstruction of the $495.4nm$ channel, which underwent reconstruction, markedly improving the MTF function. Figure 9(c) demonstrates the successful reconstruction of the image within the yellow box, revealing clear textures in each band and restoring high-frequency details from aliased data.

Tradeoff between accuracy and spectral resolution. In ADIS, spectral resolution hinges on dispersion distance,

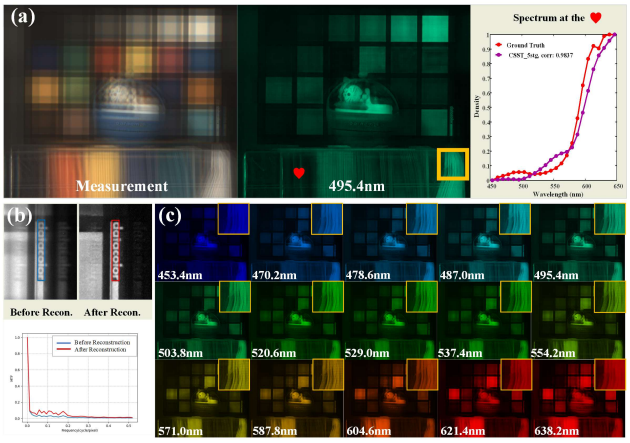

Figure 9. (a) Measurement and reconstruction results of a spectral-interesting, texture-complex scene, with a comparison of reconstructed spectra and ground truth spectra at the heart-shaped markers; (b) MTF comparison of the images before and after reconstruction; (c) reconstruction results of the scene in various bands.

while reconstruction accuracy is related to PSF concentration. A higher PSF dispersion decreases inter-band spectral data correlation, thereby alleviating underdetermination in the inverse process. Hence, future efforts should center on optimizing system parameters and algorithm performance to enhance overall performance.

Sparse Propensity of Reconstruction. Comparing the reconstruction results of different scenes in Figure 7(b) and 9(c), the artifacts within ADIS reconstructions escalate when the texture complexity and spectral complexity intensify, which could potentially be mitigated through augmentation of training data complexity and diversity.

6. Conclusion

A compact diffractive optical system comprising an ultra-thin aperture mask and conventional imaging lens forms a discrete coding pattern on a mosaic sensor. The Cascaded Shift-Shuffle Spectral Transformer (CSST) algorithm is used to decode the diffraction pattern for high-resolution hyperspectral imaging. Meanwhile, the system's spatial invariance ensures pattern robustness, and its diffraction efficiency is improved to 75% using Babinet's principle. Further work is needed to improve imaging quality and spectral resolution while maintaining high diffraction efficiency. Furthermore, there's a need to investigate ADIS's potential for fulfilling large FOV demands.

7. Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62025108), the Leading Technology of Jiangsu Basic Research Plan (No.BK20192003), and the Key & Plan of Jiangsu Province (No. BE2022155).

References

[1] Xun Cao. Hyperspectral/multispectral imaging. In *Computer Vision: A Reference Guide*, pages 592–598. Springer, 2021.

[2] Quan Yuan, Qin Ge, Linsen Chen, Yi Zhang, Yuhang Yang, Xun Cao, Shuming Wang, Shining Zhu, and Zhenlin Wang. Recent advanced applications of metasurfaces in multi-dimensions. *Nanophotonics*, (0), 2023.

[3] 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.

[4] Michael Descour and Eustace Dereniak. Computed-tomography imaging spectrometer: experimental calibration and reconstruction results. *Applied optics*, 34(22):4817–4826, 1995.

[5] Xun Cao, Hao Du, Xin Tong, Qionghai Dai, and Stephen Lin. A prism-mask system for multispectral video acquisition. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2423–2435, 2011.

[6] Qionghai Dai, Chenguang Ma, Jinli Suo, and Xun Cao. Computational hyperspectral imaging. In *JSAP Annual Meetings Extended Abstracts The 75th JSAP Autumn Meeting 2014*, pages 3821–3821. The Japan Society of Applied Physics, 2014.

[7] Jacek Hunicz and Dariusz Piernikarski. Investigation of combustion in a gasoline engine using spectrophotometric methods. In *Optoelectronic and Electronic Sensors IV*, volume 4516, pages 307–314. SPIE, 2001.

[8] Adrian Taruttis and Vasilis Ntziachristos. Advances in real-time multispectral optoacoustic imaging and its applications. *Nature photonics*, 9(4):219–227, 2015.

[9] Nathan Hagen. Survey of autonomous gas leak detection and quantification with snapshot infrared spectral imaging. *Journal of Optics*, 22(10):103001, 2020.

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

[11] Xia Hua, Yujie Wang, Shuming Wang, Xiujuan Zou, You Zhou, Lin Li, Feng Yan, Xun Cao, Shumin Xiao, Din Ping Tsai, et al. Ultra-compact snapshot spectral light-field imaging. *Nature communications*, 13(1):2732, 2022.

[12] Pierre-Jean Lapray, Xingbo Wang, Jean-Baptiste Thomas, and Pierre Gouton. Multispectral filter arrays: Recent advances and practical implementation. *Sensors*, 14(11):21626–21659, 2014.

[13] Xun Cao, Tao Yue, Xing Lin, Stephen Lin, Xin Yuan, Qionghai Dai, Lawrence Carin, and David J Brady. Computational snapshot multispectral cameras: Toward dynamic capture of the spectral world. *IEEE Signal Processing Magazine*, 33(5):95–108, 2016.

[14] Michael E Gehm, Renu John, David J Brady, Rebecca M Willett, and Timothy J Schulz. Single-shot compressive spectral imaging with a dual-disperser architecture. *Optics express*, 15(21):14013–14027, 2007.

[15] Xun Cao, Xin Tong, Qionghai Dai, and Stephen Lin. High resolution multispectral video capture with a hybrid camera system. In *CVPR 2011*, pages 297–304. IEEE, 2011.

[16] Xing Lin, Gordon Wetzstein, Yebin Liu, and Qionghai Dai. Dual-coded compressive hyperspectral imaging. *Optics letters*, 39(7):2044–2047, 2014.

[17] 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.

[18] Claudia V Correa, Henry Arguello, and Gonzalo R Arce. Compressive spectral imaging with colored-patterned detectors. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7789–7793. IEEE, 2014.

[19] 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.

[20] Daniel S Jeon, Seung-Hwan Baek, Shinyoung Yi, Qiang Fu, Xiong Dun, Wolfgang Heidrich, and Min H Kim. Compact snapshot hyperspectral imaging with diffracted rotation. 2019.

[21] Sofiane Mihoubi, Olivier Losson, Benjamin Mathon, and Ludovic Macaire. Multispectral demosaicing using pseudo-panchromatic image. *IEEE Transactions on Computational Imaging*, 3(4):982–995, 2017.

[22] Shao-Wei Wang, Changsheng Xia, Xiaoshuang Chen, Wei Lu, Ming Li, Haiqian Wang, Weibo Zheng, and Tao Zhang. Concept of a high-resolution miniature spectrometer using an integrated filter array. *Optics letters*, 32(6):632–634, 2007.

[23] Nadia K Pervez, Warren Cheng, Zhang Jia, Marshall P Cox, Hassan M Edrees, and Ioannis Kymissis. Photonic crystal spectrometer. *Optics express*, 18(8):8277–8285, 2010.

[24] Andreas Tittl, Aleksandrs Leitis, Mingkai Liu, Filiz Yesilkoy, Duk-Yong Choi, Dragomir N Neshev, Yuri S Kivshar, and Hatice Altug. Imaging-based molecular barcoding with pixelated dielectric metasurfaces. *Science*, 360(6393):1105–1109, 2018.

[25] 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.

[26] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[27] Ziyi Meng, Mu Qiao, Jiawei Ma, Zhenming Yu, Kun Xu, and Xin Yuan. Snapshot multispectral endomicroscopy. *Optics Letters*, 45(14):3897–3900, 2020.

[28] Ziyi Meng, Jiawei Ma, and Xin Yuan. End-to-end low cost compressive spectral imaging with spatial-spectral self-attention. In *Computer Vision–ECCV 2020: 16th European*

Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16, pages 187–204. Springer, 2020.

[29] 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.

[30] 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.

[31] 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.

[32] Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Coarse-to-fine sparse transformer for hyperspectral image reconstruction. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 686–704. Springer, 2022.

[33] Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc Van Gool. Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging. *arXiv preprint arXiv:2205.10102*, 2022.

[34] Lizhi Wang, Chen Sun, Maoqing Zhang, Ying Fu, and Hua Huang. Dnu: Deep non-local unrolling for computational spectral imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1661–1671, 2020.

[35] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022.

[36] Z Liu, Y Lin, Y Cao, H Han, Y Wei, Z Zhang, S Lin, and B Guo. Hierarchical vision transformer using shifted windows. 2021.

[37] Guangting Wang, Yucheng Zhao, Chuanxin Tang, Chong Luo, and Wenjun Zeng. When shift operation meets vision transformer: An extremely simple alternative to attention mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2423–2430, 2022.

[38] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018.

[39] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 492–511. Springer, 2020.

[40] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14821–14831, 2021.

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

[42] 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.

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

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

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

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

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

[48] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.