Learned Multi-aperture Color-coded Optics for Snapshot Hyperspectral Imaging

ZHENG SHI*, Princeton University, USA
XIONG DUN*, Tongji University, China
HAOYU WEI, The University of Hong Kong, China
SHIYU DONG, ZHANSHAN WANG, XINBIN CHENG, Tongji University, China
FELIX HEIDE, Princeton University, USA
YIFAN PENG, The University of Hong Kong, China

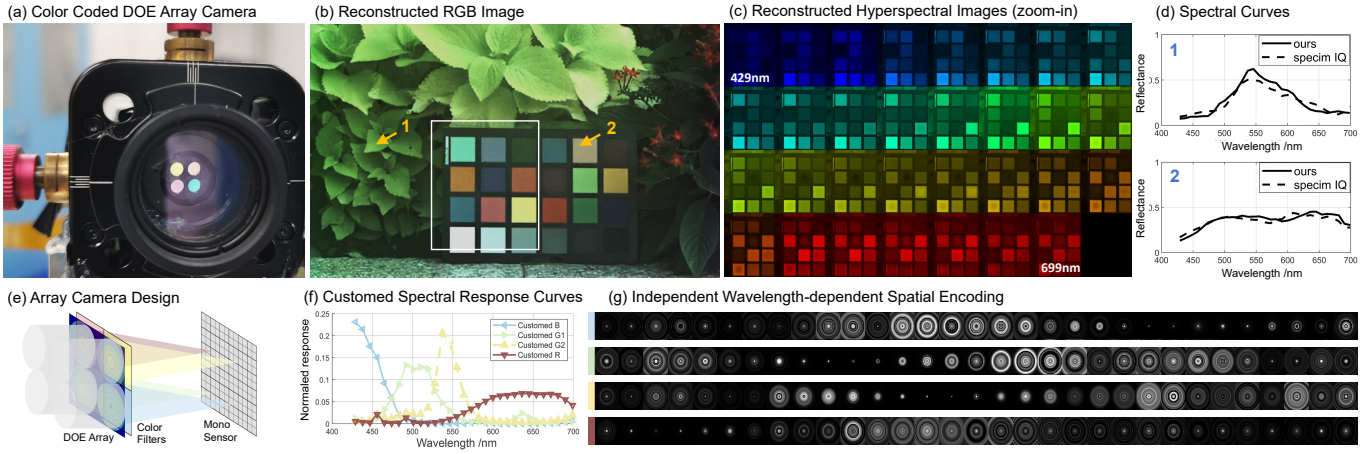

Fig. 1. We propose a snapshot hyperspectral imager with multi-aperture color-coded optics, illustrated in (a/e), providing customized independent spatial and spectral encoding for different channels, as shown in (f/g). We achieve this by jointly optimizing a DOE array, aperture-wise color filters, and a reconstruction network. This approach exploits the degrees of freedom in optical encoding across both spatial and spectral dimensions, outperforming existing single-lens approaches by over 5 dB PSNR in reconstruction quality. We experimentally validate the proposed method in both indoor and outdoor settings, recovering up to 31 spectral bands within the 429–700 nm range, closely matching the reference captured by a spectral-scan hyperspectral camera, as shown in (b-d).

Learned optics, which incorporate lightweight diffractive optics, coded-aperture modulation, and specialized image-processing neural networks, have recently garnered attention in the field of snapshot hyperspectral imaging (HSI). While conventional methods typically rely on a single lens element paired with an off-the-shelf color sensor, these setups, despite their widespread availability, present inherent limitations. First, the Bayer sensor's spectral response curves are not optimized for HSI applications, limiting spectral fidelity of the reconstruction. Second, single lens designs rely on a single diffractive optical element (DOE) to simultaneously encode spectral information and maintain spatial resolution across all wavelengths, which constrains spectral encoding capabilities. This work investigates a multi-channel lens array combined with aperture-wise color filters, all co-optimized alongside an image reconstruction network. This configuration enables independent spatial encoding and spectral response for each channel, improving optical encoding across both spatial and spectral dimensions. Specifically, we validate that the method achieves over a 5dB improvement in PSNR for spectral reconstruction compared to existing single-diffractive lens and coded-aperture techniques. Experimental validation further confirmed that the method is capable of recovering up to 31 spectral bands within the 429–700 nm range in diverse indoor and outdoor environments.

CCS Concepts: • **Computing methodologies** → **Computational photography**.

Additional Key Words and Phrases: Computational Imaging, Co-Designed Optics, Hyperspectral Imaging.

ACM Reference Format:
Zheng Shi, Xiong Dun, Haoyu Wei, Shiyu Dong, Zhanshan Wang, Xinbin Cheng, Felix Heide, and Yifan Peng. 2024. Learned Multi-aperture Color-coded Optics for Snapshot Hyperspectral Imaging . *ACM Trans. Graph.* 43, 6, Article 208 (December 2024), 11 pages. https://doi.org/10.1145/3687976

*denotes equal contribution. Tongji University is the first affiliation.
Corresponding authors: evanpeng@hku.hk, dunx@tongji.edu.cn.

Authors' Contact Information: Zheng Shi, zhengshi@princeton.edu, Princeton University, USA; Xiong Dun, dunx@tongji.edu.cn, Tongji University, China; Haoyu Wei, haoyuwei@connect.hku.hk, The University of Hong Kong, China; Shiyu Dong, Zhanshan Wang, Xinbin Cheng, dongsy@tongji.edu.cn,wangzs@tongji.edu.cn,chengxb@tongji.edu.cn, Tongji University, China; Felix Heide, fheide@princeton.edu, Princeton University, USA; Yifan Peng, evanpeng@hku.hk, The University of Hong Kong, China.

ACM 1557-7368/2024/12-ART208
https://doi.org/10.1145/3687976

1 Introduction

Hyperspectral imaging (HSI) is a high-dimensional data acquisition process that yields a series of 2D images of a single scene across finely resolved wavelength bands [Zhang et al. 2023b]. This imaging modality holds promise across diverse domains, such as medical imaging [Lu and Fei 2014], remote sensing [Hinderberger et al. 2023], agriculture monitoring [Lu et al. 2020], pattern recognition [Rao et al. 2022], and industrial assessment [Liu et al. 2017]. Existing methods rely on scanning-type acquisition systems, often constrained by extended exposure time and/or intricate hardware components [Cao et al. 2016], thereby prohibiting real-time imaging and sensing applications.

Snapshot HSI methods, based on the principle of compressed sensing (CS), significantly reduce the acquisition load of either spatial or spectral information through coded modulation [Arguello et al. 2021]. The past decade has witnessed a number of refractive optics-based snapshot HSI devices, for example, the coded-aperture spectral imager [Li et al. 2012], the spatial-spectral and/or diffuser encoded compressive imaging [Monakhova et al. 2020], the mask-guided spectral-wise image reconstruction [Cai et al. 2022a], the optimized broadband filter encoding [Zhang et al. 2021], and the single-pixel camera spectrometer [August et al. 2013].

Recent research increasingly focuses on utilizing compact diffractive optical elements (DOEs) as thin lenses, which offer unparalleled design flexibility at the micro scale [Dun et al. 2020; Jeon et al. 2019; Peng et al. 2016; Sitzmann et al. 2018; Xu et al. 2023]. Significant advancements have also been made in end-to-end (E2E) designed diffractive optical systems that leverage machine intelligence [Meng et al. 2021; Tseng et al. 2021; Zhang et al. 2022, 2021]. However, existing E2E-designed spectral imaging systems typically do not include customized color filters with single-diffractive lenses, largely due to the complexities involved in manufacturing custom spectral response curves. Such customization often requires extensive lithography and coating processes [Dong et al. 2018; Yako et al. 2023], which can be costly and time-consuming. Moreover, the encoding capabilities of a single-aperture element, especially when paired with a conventional Bayer-type color filter array (CFA), are significantly constrained, further limiting their practicality in high-fidelity applications.

To address these limitations, we further the E2E optimization paradigm by incorporating color filters into the optical design, and introduce an alternative multi-aperture approach. We employ aperture-wise color filters combined with a multi-aperture diffractive lens element array, rather than traditional single DOE and pre-designed Bayer CFAs combination. This configuration not only potentially simplifies the manufacturing process but also significantly improves spectral encoding capability by allowing for independent spatial and spectral encoding across different color channels. This enhanced encoding capability captures a richer spectrum of information, which, when processed by a co-optimized reconstruction network, leads to more accurate spectral reconstructions. We assess the proposed approach both in simulation and with an experimental prototype, validating that our method can accurately recover up to 31 spectral bands within the 429–700 nm range under various lighting conditions.

Specifically, we make the following contributions:

- We introduce a multi-aperture snapshot hyperspectral imaging system that provides independent spatial and spectral encoding across individual channels. Our system features a 2×2 array configuration, integrating diffractive lenses and aperture-wise customized color filters with a monochrome sensor, all co-optimized with a reconstruction network to enhance performance.
- We analyze the method and compare it with existing single-diffractive lens and coded-aperture solutions, demonstrating an improvement of over 5 dB in PSNR across 31 spectral bands (429–700 nm) for snapshot hyperspectral reconstruction.
- We develop and experimentally validate a prototype system against a Specim IQ scanning hyperspectral camera, confirming that the proposed method is capable of accurately resolving 31 spectral bands in diverse indoor and outdoor scenes.

Lens designs, color filter designs, network checkpoints, and all code necessary to reproduce the results are available at the authors' webpage.

Overview of Limitations. Our prototype system is constructed in an academic facility, where the fabrication quality of optical elements is considerably lower than that achieved by state-of-the-art processes. This discrepancy between fabrication and design leads to noticeable diffraction efficiency loss and haze artifacts in the captured images. While our multi-aperture approach is not confined to the 2×2 configuration, we have selected this setup based on a trade-off where image resolution is compromised because the full sensor measurement area is divided among multiple channels. Specifically, our prototype sensor with a resolution of 2,048×2,048 enables a reconstruction resolution of 1,024×1,024. Utilizing a higher-resolution image sensor could potentially enhance spectral reconstruction performance without sacrificing spatial resolution.

2 Related Work

Computational Cameras with Differentiable Diffractive Optics. Traditional imaging systems use compound refractive lenses engineered for perceptual quality, focusing on color balance and sharpness [Malacara et al. 2003]. However, these systems often struggle with specialized vision tasks like seeing through occlusions [Shi et al. 2022] or depth estimation [Li et al. 2022]. To address these gaps, extensive research in computational photography has led to the development of specialized lens systems employing diffractive optical elements (DOEs). DOEs, with their micron-scale profiles, enable precise modulation of light's phase through diffraction [Levin et al. 2007]. Such systems can be optimized using back-propagation [Sitzmann et al. 2018; Wang et al. 2022], modeling the image formation process with differentiable wave optics. When combined with learnable reconstruction algorithms, these diffractive optics not only support high-quality color imaging [Peng et al. 2019] but also facilitate advancements in microscopy [Liu et al. 2022b; Nehme et al. 2020], hyperspectral imaging [Baek et al. 2021; Jeon et al. 2019; Li et al. 2022], super-resolution and extended depth of field [Sun et al. 2021].

Multi-Aperture Cameras. Researchers have long explored capturing multiple images simultaneously with a single camera setup. For example, Green et al. [2007] introduced a design that splits the aperture into a central disc to capture 2×2 images on the sensor with different aperture sizes. In another stream of research [Brückner et al. 2010; Chakravarthula et al. 2023; Tanida et al. 2001], inspired by the compound eyes of insects, has developed miniature cameras that overcome the trade-offs between focal length and field of view using arrays of thin lenses. In the realm of color imaging, Venkataraman et al. [2013] addressed chromatic aberrations using an on-sensor array of color-filtered single lens elements, transforming the deconvolution challenge into a chromatic light field reconstruction task. Building on these developments, our method integrates DOEs with aperture-wise color filters to achieve customized, independent spatial and spectral encoding for each cholor channel, enhancing snapshot hyperspectral imaging capabilities.

Snapshot Hyperspectral Imaging. To circumvent the extended exposure times required by traditional scanning-type hyperspectral systems, researchers have explored alternative snapshot approaches. Earlier methods relied on bulky setups involving multiple optical components such as dispersive elements (prisms), coded apertures, and several relay and imaging lenses, which made them impractical for many applications [Baek et al. 2017]. In pursuit of compact snapshot spectral imaging systems, more recent approaches have employed diffractive optical elements (DOEs) with spectrally varying point spread functions to encode hyperspectral information [Baek et al. 2021; Jeon et al. 2019; Li et al. 2022; Xu et al. 2023]. To simultaneously sample the angular and spectral dimensions, Xiong et al. [2017] used beam splitters to integrate an off-the-shelf light field camera with a coded-aperture snapshot spectral imager (CASSI) into a single setup. Additionally, some studies have investigated diffractive neural networks to resolve multi-spectral images without spectral filters [Mengu et al. 2023], though these face challenges related to assembly complexity and light efficiency. There is also ongoing research into nano-fabricated optics, such as meta optics [Hua et al. 2022; Lin et al. 2023; Zhang et al. 2023a], but their viability for consumer products and overall imaging quality is still underdeveloped. In this work, we introduce a multi-aperture setup that enhances the design flexibility and encoding capabilities of diffractive optical systems, potentially overcoming the limitations of previous designs and pushing the boundaries of spectral imaging technology.

3 Multi-Aperture Color-Coded Hyperspectral Camera

The proposed hyperspectral imaging (HSI) system is composed of an optical and a computational module. The optical module includes a diffractive multi-lens array, an array of aperture-wise color filters, and a monochrome sensor that captures the scene through multiple channels with separate encoding. Following this, a co-designed deep neural network computationally reconstructs hyperspectral images from the sensor measurements. Figure 2 provides an illustration of the multi-aperture color-coded snapshot hyperspectral camera. In this section, we describe the differentiable forward imaging model, the reconstruction network, and the hybrid loss function that enables the co-optimization of this computational camera.

3.1 Forward Imaging Model with Multi-aperture Optics

We start by outlining the multi-aperture image formation model that facilitates the joint optimization of the lens array, aperture-wise CFA, and the reconstruction network. While the proposed multi-aperture approach can accommodate various array sizes, for simplicity, we will focus on an optical setup composed of a 2×2 sub-aperture array in our examples.

The incident wave field of each sub-DOE unit is modulated by the corresponding color filter placed on the sub-lens' aperture. Given that the transmission curve of the color filter is $T_c(\lambda)$, where footnotes $c = 1, 2, 3, 4$ indicate the color channels, images $Y_c(x, y)$ acquired by the DOE, color filter, and sensor can be formulated as

$$Y_c(x, y) = \int_{\lambda_{min}}^{\lambda_{max}} [PSF_c(x, y, \lambda) * I(x, y, \lambda)] T_c(\lambda) T_s(\lambda) d\lambda, \quad (1)$$

where $[\lambda_{min}, \lambda_{max}]$ indicates the spectrum range, and $T_s(\lambda)$ represents the monochromatic sensor's native transmission curve. The PSF_c for each sub-DOE is formulated with the rotational symmetric diffractive lens representation [Dun et al. 2020] as

$$PSF_c(\rho, \lambda) = \left| \frac{2\pi}{\lambda f} e^{i\frac{k}{2f}(\lambda f \rho)^2} \sum_{m=1}^{\infty} P(r_m, \lambda) e^{\frac{ik}{2f} r_m^2} H(r_m, \rho) \right|^2, \quad (2)$$

where f represents the distance between the DOE and the sensor, equivalent to the focal length of all sub-lenses. The complex transmittance function of the DOE is denoted as

$$P(r_m, \lambda) = A(s, t) e^{ik(n(\lambda)-1)h(s,t))}. \quad (3)$$

Additionally, ρ is defined as $\rho = \frac{\sqrt{x^2+y^2}}{\lambda f}$, k is the wave number $k = 2\pi/\lambda$, and $n(\lambda)$ is the refractive index of the substrate. The height map and aperture of each sub-DOE unit are represented by $h(s, t)$ and a *circ* function $A(s, t)$, respectively, while spatial coordinates at the DOE and sensor planes are denoted as (s, t) and (x, y). The function $H(r_m, \rho)$ is defined as

$$H(r_m, \rho) = \begin{cases} \frac{1}{2\pi\rho}[r_m J_1(2\pi\rho r_1) - r_{m-1} J_1(2\pi\rho r_{m-1})], & m > 1 \\ \frac{1}{2\pi\rho} r_1 J_1(2\pi\rho r_1), & m = 1 \end{cases}, \quad (4)$$

where J_1 is the 1^{st} order Bessel function of the first kind.

By modeling the sensor noise as a pixel-wise Gaussian-Poisson noise, the sensor output can be formulated as

$$S_c(x, y) = \eta_p(Y_c(x, y), \sigma_p) + \eta_g(Y_c(x, y), \sigma_g), \quad (5)$$

where $\eta_g(Y_c(x, y), \sigma_g) \sim N(Y_c(x, y), \sigma_g^2)$ is the Gaussian noise component, and $\eta_p(Y_c(x, y), \sigma_p) \sim P(Y_c(x, y)/\sigma_p)$ is the Poisson noise component. Eventually, four encoded, grayscale images $S_c(x, y)$s are obtained and then input to the reconstruction network.

3.2 Image Reconstruction Neural Network

The proposed recovery network architecture comprises two sequential components: a feature-extraction block responsible for multi-scale feature extraction from the sensor measurements, and a pair of reconstruction heads that individually resolve either RGB or HSI output based on the extracted features.

Inspired by recent advancements in image classification networks [Wang et al. 2020], we incorporate parallel convolution streams

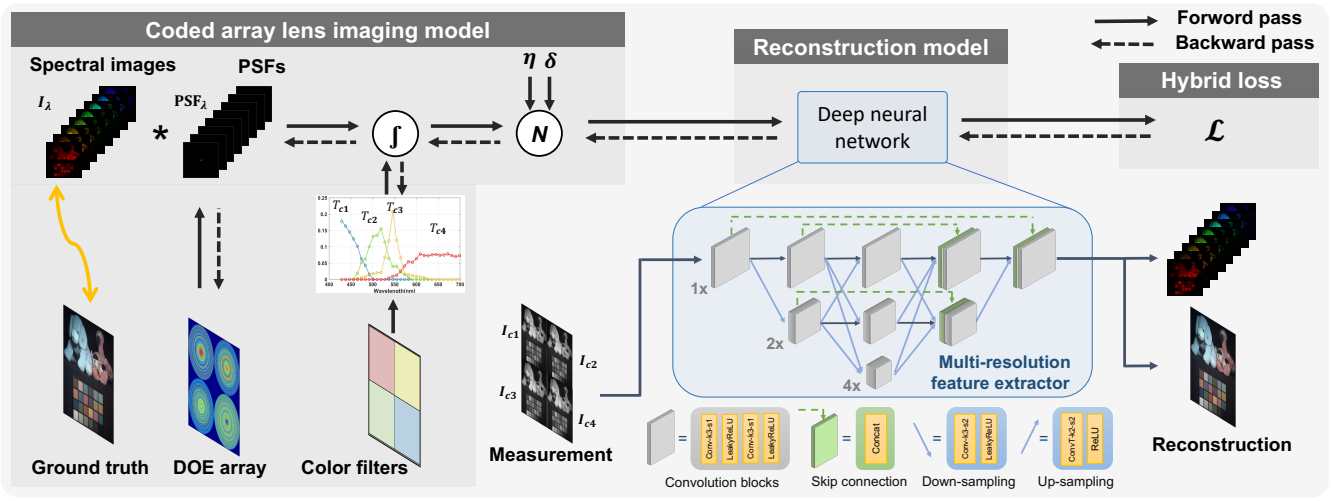

Fig. 2. **Learning Multi-Aperture Color-Coded Optics for Snapshot Hyperspectral Imaging.** We jointly optimize the multi-aperture DOE array, aperture-wise color filters, and image reconstruction network using a hybrid loss function. During each forward pass, the ground truth spectral images are first convolved with the PSFs of the DOE array and then multiplied by the response curves of the color filters. Noise is added to the simulated sensor image, which is then integrated over the monochrome sensor's response for each sub-lens channel: B, G1, G2, and R. These images are input into the multi-resolution feature extractor of the image reconstruction network to recover the final hyperspectral (HS) and RGB images.

at three different resolution levels (1x, 2x, and 4x down-sampling) within the entire feature-extraction block to preserve high-resolution features. Additionally, concatenation layers are utilized to aggregate features from different resolutions, enabling information flow between streams; and skip connections are employed to retain all texture details captured by sensor measurement. Our reconstruction heads are convolution blocks with 5×5 kernels, which consume the output of the feature-extraction block and are specialized in RGB or HSI reconstruction. A detailed architecture spreadsheet is presented in the supplemental document Tbl. S2.

3.3 Hybrid Loss with Fabrication-aware Regularization

The proposed pipeline, including both the optical component and the reconstruction network, is trained by minimizing a hybrid loss function

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \mathcal{R}_{\text{PSF}} + \mathcal{R}_{\text{T}}. \tag{6}$$

This loss function is comprised of three key components: an end-to-end reconstruction loss, denoted as $\mathcal{L}_{\text{recon}}$, which evaluates the quality of the reconstruction, as well as two regularization terms, $\mathcal{R}_{\text{PSF}}$ and $\mathcal{R}_{\text{T}}$, which account for encouraging fabrication-friendly DOE and color filter designs, respectively.

The reconstruction loss $\mathcal{L}_{\text{recon}}$ can be further broken down as

$$\mathcal{L}_{\text{recon}} = w_1 \mathcal{L}_{\ell_1}^{\text{RGB}} + w_2 \mathcal{L}_{\text{perc}}^{\text{RGB}} + w_3 \mathcal{L}_{\ell_1}^{\text{Spectral}}. \tag{7}$$

Here, $\mathcal{L}_{\ell_1}$ represents a pixel-wise ℓ_1 loss, quantifying the deviation between the reconstructed RGB and spectral images and their corresponding targets. Additionally, we incorporate $\mathcal{L}_{\text{perc}}$, a perceptual loss based upon LPIPS [Zhang et al. 2018], which is applied to the reconstructed RGB images to capture perceptual dissimilarities. w_1, w_2, w_3 denote the weights for each loss term, empirically set at 100,1 and 100.

$\mathcal{R}_{\text{PSF}}$ looks at the intensity at the center of the learned PSFs and discourages overly blurry PSFs, that is

$$\mathcal{R}_{\text{PSF}} = \begin{cases} 0, & \text{if } P_{\text{center}} \geq P_{\text{target}} \\ P_{\text{target}} - P_{\text{center}}, & \text{if } P_{\text{center}} < P_{\text{target}}. \end{cases} \tag{8}$$

Here, P_{center} represents the total intensity of a 30×30 center crop of the learned PSF, while P_{target}, empirically set at 0.9, represents the target total intensity.

Finally, $\mathcal{R}_{\text{T}}$ is employed to induce a smooth color filter design and discourage multi-modal curves, that is

$$\mathcal{R}_{\text{T}} = \max_{i=1}^{n-1} |T_c(\lambda_{i+1}) - T_c(\lambda_i)| + \sum_{i=1}^{n-1} |T_c(\lambda_{i+1}) - T_c(\lambda_i)|. \tag{9}$$

This regularization term computes the 1st order derivatives of the learned color filter T_c. It encourages that the color filter design maintains smoothness and helps prevent local maxima, both of which facilitates fabrication.

As illustrated in Fig. 2, the proposed loss function is directly applied to the reconstructed network output, so as to supervise both optics optimization, including DOEs and color filters, and image reconstruction network update in an end-to-end manner.

4 Analysis

Before evaluating our method with experimental measurements, we first validate the method on synthetic data. We begin by comparing our method to existing snapshot HSI approaches in Section 4.2 to verify its effectiveness. Next, we assess the benefit of our multi-aperture setup in Section 4.3, followed by an ablation study of the chosen hybrid loss function and the architecture of the reconstruction network in Section 4.4.

4.1 Datasets and Training Details

To ensure the generalization ability of the proposed method, we utilize two datasets: CZ_HSDB [Chakrabarti and Zickler 2011] and ICVL [Arad and Ben-Shahar 2016] (278 images in total), for training purposes, while evaluating the performance of the method on the previously unseen CAVE [Yasuma et al. 2008] dataset (32 images) and KAUST [Li et al. 2021] dataset (409 images, quantitative performance reported in the supplemental document Sec. 4). To account for variations in image sizes across these datasets, we randomly extract 512×512 pixels' crops of the images during training. In addition, left-right flips and random channel shuffling are applied to augment the data during the training process.

4.2 Synthetic Assessment

In the following, we assess the performance of our proposed method by comparing it to existing snapshot HSI approaches. We consider two categories of baseline methods: (1) Direct spectral reconstruction from RGB Images, represented by the HRNet method [Zhao et al. 2020] and MST++ method [Cai et al. 2022b], both are NTIRE Spectral Reconstruction Challenge Winners; and (2) Alternative compressive snapshot spectral imaging systems, represented by QDO [Li et al. 2022] and SCCD [Arguello et al. 2021], both of which are state-of-the-art E2E optimized diffractive optics-based snapshot imaging systems.

We provide qualitative and quantitative comparisons in Fig. 3 and Tab. 1, respectively. Additional comparisons are shown in Fig. S6 (left, rows 2-4) in the supplemental document. For these comparisons, we feed the baseline model with either the ground-truth RGB images with simulated noise, or simulated sensor measurements using the optics design and specs provided by the competing baseline methods. All inputs are of size $512 \times 512 \times 3$. We follow the forward model described in Section 3.1 when needed and use the same RGB simulation curve for the HRNet method and MST++ method. Since both QEO and SCCD methods optimized the optics design and reconstruction in an end-to-end manner, we utilize pre-trained reconstruction models provided by the authors of these works. Note, while SCCD can reconstruct a total of 49 spectral bands spanning the 420 – 660nm range, only 25 bands within this range have been configured in the provided network to recover, with intervals of 10 nm between each. Consequently, this results in the absence of certain spectral channels, as illustrated in Fig. 3. In addition to conventional image metrics SSIM and PSNR, we employ two domain-specific metrics to more comprehensively evaluate spectral reconstruction performance. In the RGB domain, we use Delta E [Sharma et al. 2005], which quantifies the perceptual difference between the ground truth and the reconstruction, with Delta E ≤ 1 being considered imperceptible to the human eye. In the hyperspectral domain, we use the Spectral Angle Mapper (SAM) [Kuching 2007], which computes the spectral angle between the reconstructed and ground truth spectra in an n-dimensional spectral space, where n represents the number of spectral bands.

As shown in Fig. 3 and Tab. 1, the QDO [Li et al. 2022] method exhibits the poorest reconstruction performance, primarily due to its reliance on a single-lens design and the use of heavy quantization during the design phase. The quantization, intended to align the DOE design with its actual fabrication, significantly limits design flexibility. On the other hand, SCCD [Arguello et al. 2021], which incorporates a jointly optimized color-coded aperture (CCA) for spatial and spectral modulation, achieves a 3.8 dB improvement in spectral reconstruction accuracy. However, its constrained spectral filter coding and dependence on a pre-designed Bayer-patterned sensor leads to a noticeable gap disparity between its reconstructions and the ground truth. In contrast, our proposed method, leveraging customized spectral coding and advanced coding capabilities through an array design, achieves a substantial 5.4 dB improvement in PSNR for HS images and a 5.7 dB enhancement for RGB images.

Table 1. **Quantitative comparison over the 32 unseen images in CAVE dataset.** We compare our reconstruction network architecture with state-of-the-arts, including RGB-to-Spectrum reconstruction represented by HRNet [Zhao et al. 2020] and MST++ [Cai et al. 2022b], and recent compressive snapshot spectral imaging systems, represented by QDO [Li et al. 2022] and SCCD [Arguello et al. 2021]. Note, that we do not report the RGB output scores for HRNet and MST++ as it takes RGB as input, and the SCCD scores are not punished for the missing channels.

	RGB			HS		
	SSIM	PSNR	Delta E	SSIM	PSNR	SAM ↓
Proposed	**0.96**	**38.01**	**1.30**	**0.92**	**32.82**	**0.21**
HRNet [2020]	-	-	-	0.88	26.76	0.40
MST++ [2022b]	-	-	-	0.833	24.85	0.41
SCCD [2021]	0.84	32.27	2.53	0.74	27.43	0.59
QDO [2022]	0.74	26.19	4.32	0.67	23.60	0.45

Diverging from these optical encoding techniques, HRNet [Zhao et al. 2020] and MST++ [Cai et al. 2022b] are state-of-the-art RGB-to-HS methods that directly reconstructs hyperspectral information from conventional RGB captures. Although they circumvents the inherent spatial resolution loss linked to DOE spectral coding and provides high-quality plausible outputs, it is important to recognize that RGB-to-HS approaches face the ill-posed challenge of the inverse problem, and can only attempt to extrapolate the missing spectral information based on the learned image prior from the training dataset, making them susceptible to overfitting. As a result, MST++ is here overfitting to the specific capture setup (e.g., JPEG compression and camera calibration), leading to a 7.97 dB drop in reconstruction quality (PSNR) compared to the proposed method. While HRNet achieves better reconstruction quality, as it is designed to perform well with unknown and uncalibrated cameras [Arad et al. 2020], it still shows a 6.04 dB (PSNR) reduction compared to the proposed method.

4.3 Analysis of Multi-Aperture Configuration

Next, we validate the proposed multi-aperture setup by examining the performance enhancements from spatial and spectral modulation. We also assess the impact of supporting independent spatial modulation across each channel versus shared modulation across channels. To ensure a fair comparison and avoid influences from other components such as training procedures and sensor settings, each setting produces $512 \times 512 \times 4$ sensor outputs, which are then passed through the same network architecture trained using identical configurations. Quantitative results are reported in Tab. 2, and

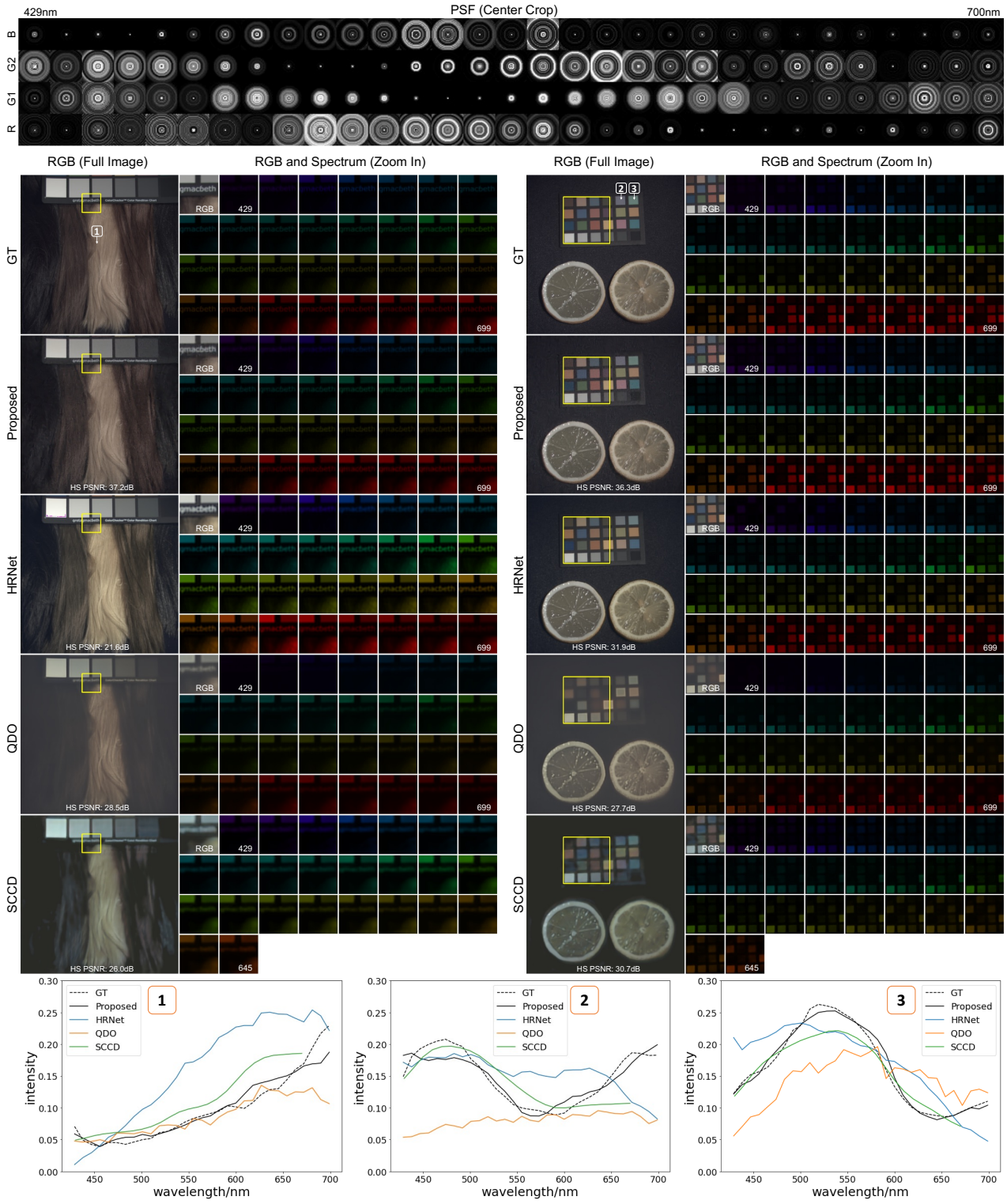

Fig. 3. **Assessment in Simulation.** We compare the ground truth spectra (GT, 1st row) to reconstructions from our proposed method (2nd row), the RGB-to-Spectrum method HRNet [Zhao et al. 2020] (3rd row), and recent compressive snapshot spectral imaging systems, namely QDO [Li et al. 2022] (4th row) and SCCD [Arguello et al. 2021] (5th row). For each subset, we show the RGB recovery results of full images on the left and the zoom-in version of both RGB and spectrum recovery results on the right. We also present spectral validation plots of all approaches for three specific points, labelled 1, 2, and 3 on the 1st row RGB images, displayed at the bottom of the figure. Center-cropped designed PSFs for four channels at the target wavelengths are visualized at the top.

Table 2. **Validation of Multi-Aperture Configuration.** We assess the benefits of the proposed multi-aperture setup by analyzing the performance enhancements from spatial and spectral modulation, as well as the effects of independent versus shared spatial modulation across channels. To this end, we compare the proposed approach to variants using a fixed Bayer RGGB color filter and/or a single shared DOE across all color channels.

	RGB		HS		
	SSIM	PSNR	SSIM	PSNR	SAM ↓
Proposed	0.96	38.01	**0.92**	**32.82**	**0.21**
Thin Lens w/ Fixed Curve	**0.99**	**46.29**	0.88	30.67	0.36
Thin Lens w/ Learned Curve	0.97	44.01	0.90	30.91	0.33
Single DOE w/ Learned Curve	0.84	34.96	0.76	29.84	0.43
Single DOE w/ Fixed Curve	0.78	31.83	0.69	27.42	0.50
DOE Array w/ Fixed Curve	0.94	37.80	0.86	30.32	0.35

qualitative results are available in Fig. 4 , Fig. S4, and rows 2 to 5 (right) of Fig. S6 in the supplemental document.

We first simulate a baseline scenario with no wavelength-dependent spatial modulation, paired with a pre-designed fixed RGGB Bayer color filter for spectral encoding, represented by the 'Thin Lens w/ Fixed Curve' configuration. This setup can be considered the standard RGB-to-HS setting in an ideal case, where the scene is all-in-focus and the sensor response curve is known to the algorithm. As a result, the network is able to reconstruct a near-perfect RGB image, since it is provided as an input, and achieves good performance in the hyperspectral domain because the sensor's response curve is known to the model. However, in typical RGB-to-HS settings, the algorithms are often required to provide reasonable predictions regardless of the sensor's response curve, resulting in less accurate hyperspectral reconstruction.

By adding optimized spectral modulation via a customized color filter $T_c(\lambda)$, 'Thin lens w/ Learned Curve' configuration improves hyperspectral performance by providing different spectral encoding for the two otherwise identical G channels in a Bayer filter. However, while bypassing the spatial resolution loss typically induced by DOE modulation, both configurations must interpolate the data from a 4-channel measurement into 31 channels in the absence of spatial modulation, leading to diminished hyperspectral performance compared to the proposed method.

We next assess the impact of utilizing an array setup where each color channel has its own independent spatial modulation, compared to utilizing a single DOE that shares modulation for all four channels. To accomplish this, we optimize a single DOE with either a fixed RGGB Bayer filter or a customizable 4-channel color filter, reporting performance under the 'Single DOE w/ Fixed Curve' and 'Single DOE w/ Learned Curve' configurations. This shared modulation significantly limits design flexibility, as a single DOE must encode additional spectral information while simultaneously maintaining reasonable spatial resolution for all wavelengths. Due to the nature of diffraction, the DOE tends to focus on a narrow wavelength band, making it challenging to reconstruct high-frequency spatial details across other wavelengths. Consequently, we observe a noticeable decline in both RGB and hyperspectral reconstruction quality when compared to the proposed multi-aperture configuration.

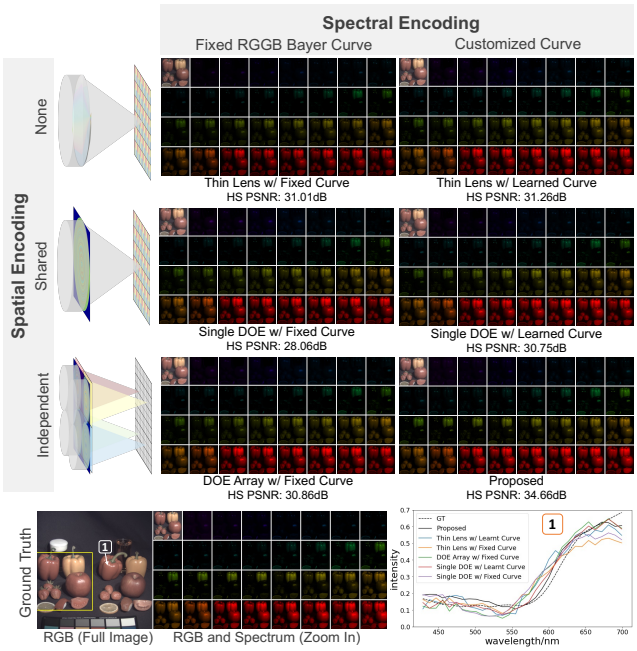

Fig. 4. **Qualitative Analysis of the Multi-Aperture Configuration.** We evaluate the benefits of the proposed multi-aperture setup by analyzing performance enhancements from spatial and spectral modulation, as well as the effects of independent versus shared spatial modulation across channels. The columns compare a fixed Bayer color filter (left) with a learned color filter (right), while the rows compare no spatial encoding (first row), shared spatial encoding (single DOE, second row), and independent spatial encoding (multi-aperture setup, third row). The ground truth scene is displayed at the bottom left, and the spatial response curve from all configurations of a sampled point is shown at the bottom right. Refer to the Fig. S4 in supplemental document for a higher-resolution error map.

To further evaluate the performance enhancement from spectral modulation provided by the customized color filter $T_c(\lambda)$, we conduct an additional experiment, 'DOE Array w/ Fixed Curve,' where a multi-aperture DOE array is paired with a fixed RGGB Bayer filter. We confirm that Bayer filters, designed to mimic human visual perception, are not tailored to HSI applications, leading to a performance decline compared to our proposed method.

4.4 Reconstruction Ablation Experiments

Next, we validate the chosen regularization terms and reconstruction network architecture with an ablation study, with results reported in Fig. 5, Tab. 3 , and Fig. S6 in the supplemental document. We also include analysis on the effect of varying color filter initializations in the supplemental document Sec. 4.

Regularization Terms. We first assess the impact of the regularization terms outlined in Sec. 3.3, presenting qualitative findings in Fig. 5. Without regularization, although the optimization yields promising reconstruction results during simulations, it often converges towards designs that are challenging to fabricate or are highly susceptible to fabrication errors.

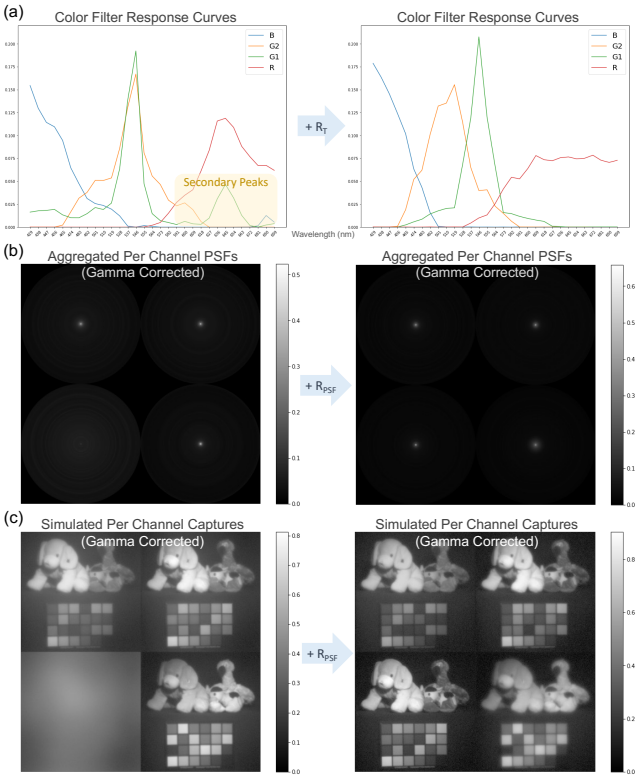

Fig. 5. **Impact of the fabrication-aware regularization terms $\mathcal{R}_{\mathbf{PSF}}$ and $\mathcal{R}_{\mathbf{T}}$.** In contrast to non-regularized designs (left), the proposed designs (right) effectively eliminate the multiple 'secondary peaks' highlighted in color filter response curves (a) and avoid excessive defocus, see the bottom left corner of aggregated per-channel PSFs (b) and simulated per-channel sensor measurements(c). Thus, $\mathcal{R}_{\text{PSF}}$ and $\mathcal{R}_{\text{T}}$ successfully improve the manufacturability of the optical system.

For instance, when omitting the regularization term $\mathcal{R}_{\text{T}}$, the learned color filters—particularly G_1 and G_2, both initialized with green Bayer filters—opted to incorporate various local maxima to capture additional spectral information, as highlighted in Fig. 5 (a). This results in designs that are difficult to fabricate. With regularization, the learned color filter response curves remain predominantly single-mode but shift their primary peak locations. Furthermore, the regularization term $\mathcal{R}_{\text{PSF}}$ prevents excessively blurry and defocusing designs, as shown in Fig. 5(b) and (c). While such designs offer enhanced spectral information in simulations, they are prone to fabrication inaccuracies and exhibit limited real-world performance.

Reconstruction Network Architecture. The assessment extends to the proposed reconstruction network architecture, wherein we evaluate the impact of different architectural components, including the multi-resolution feature extraction, skip connections, and separate reconstruction heads. As we quantitatively report in Tab. 3, the combination of the skip connection from UNet and the multi-resolution feature extraction architecture from HRNet has resulted in an improvement in both SSIM and PSNR over the vanilla UNet and HRNet-like architecture, regarding recovering both the RGB image and the hyperspectral (HS) images. Moreover, the use of separate reconstruction heads for RGB and HS images forces the network to extract HS information from wavelength-dependent defocus cues, without relying on the RGB-to-HS hallucination, which is an easier task for the network. This architectural choice has led to a significant enhancement in the quality of HS images. For qualitative comparisons, please refer to Fig. S6 (bottom 2 rows) in the supplemental document.

Table 3. **Quantitative Reconstruction Network Analysis.** We evaluate different reconstruction network architectures in simulation, including the proposed one, its variants w/o RGB recovery head and w/o skip connection, as well as the counterpart UNet.

	RGB		HS		
	SSIM	PSNR	SSIM	PSNR	SAM ↓
Proposed	**0.96**	**38.01**	**0.92**	**32.82**	**0.21**
Proposed w/o RGB head	0.95	37.42	0.86	29.90	0.32
Proposed w/o skip	0.95	36.53	0.84	29.35	0.33
Counterpart UNet	0.93	36.62	0.84	29.86	0.39

5 Experimental Assessment

To experimentally evaluate the proposed method with real-world captures, we fabricate the learned diffractive optical element and aperture-wise CFA described in Sec. 3. We first describe the experimental setup and validate that the measured PSFs and spectral curve feature the desired property, before confirming the effectiveness of the method with experimental reconstructions from our prototype camera system.

5.1 Experimental Prototype

As illustrated in Fig. 1 (a), our prototype incorporates the proposed customized DOE and color filter array, and we employ a FLIR 515M USB 3.0 monochrome sensor that offers a sensor resolution at 2,480×2,480 and a pixel pitch of 3.45 μm. The DOE array is fabricated using grayscale lithography and a molding process [Ikoma et al. 2021]. The multilayer-type color filters are designed using commodity TFCale software given the learned spectral response curves, and then fabricated through an iterative coating process [Dong et al. 2018]. For additional details on the fabrication procedures, please refer to Sec. 1 in the supplementary document .

To validate the fabrication, we measure the PSFs of each sublens and the spectral response of the corresponding aperture-wise CFA across all targeted wavelengths using an iHR 320 monochromator paired with a white light point source. The measured PSFs undergo pre-processing to approximate their design ring diameters and eliminate residual acquisition artifacts. For visualizations of the measured PSFs and spectral response curves, please refer to Fig. S2 in the supplementary document.

5.2 Experimental Results

To compensate for fabrication inaccuracies of the customize optical element, we finetune our image reconstruction neural network using the measured PSFs and spectral response curve. We validate the proposed system under both outdoor and indoor environments and compare the reconstruction with reference captures obtained from the commercially available Specim IQ hyperspectral camera.

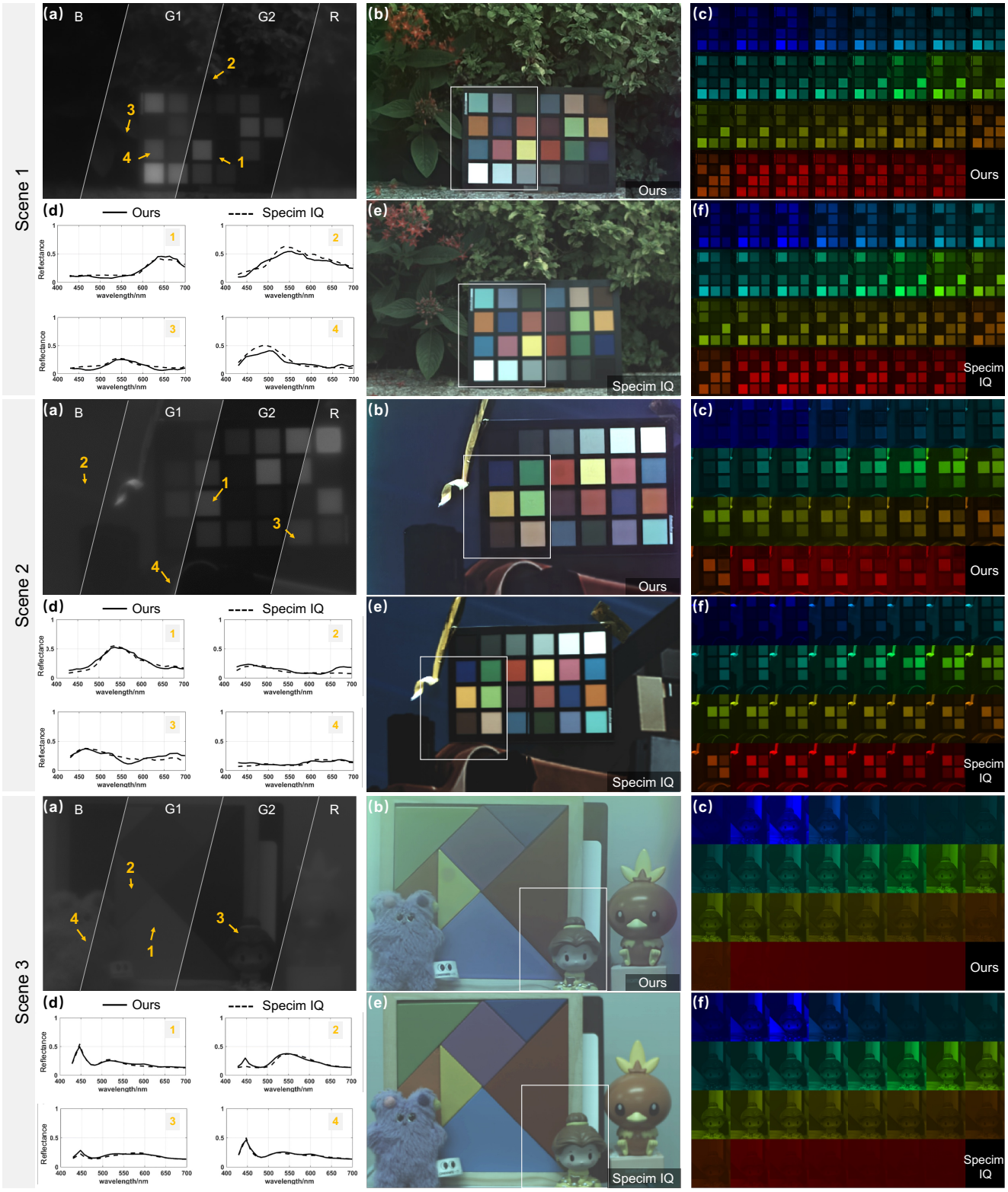

Fig. 6. **Experimental Assessment.** We evaluate the proposed method in both outdoor (Scene 1 and 2) and indoor (Scene 3) environments, comparing it to the commercial Specim IQ hyperspectral camera. For each scene, we include: (a) sensor captures comprising four sub-channel images (R, G1, G2, B); (b/e) RGB reconstructions compared to Specim IQ references; (c/f) close-up views of a cropped region across all 31 channels; and (d) spectral validation plots for four sampled points on the captured scene.

Each scene is first captured using the Specim IQ camera as a reference for qualitative evaluation. We then capture the scene with our prototype. We apply a homography-based calibration method [Peng et al. 2019; Sukthankar et al. 2001] to register the images from the four sensor channels and adjust the white balance based on the white checker in the scene. These pre-processed quad-channel images serve as input to our neural network for reconstructing HS images across 31 channels as well as a high-fidelity RGB image.

Reconstructions with both RGB and hyperspectral visualizations, along with spectral validation plots, are presented in Fig. 6, Fig. S7 and Fig. S8 in the supplemental document. To account for resolution and view angle discrepancies between the Specim IQ hyperspectral camera and our prototype, we manually selected patches of uniform color from the scene for spectral profile validation. Spectral curves reconstructed by our method closely align with those from the Specim IQ camera, demonstrating exceptional fidelity across 31 channels, further validating our methodology in diverse environments.

The integration times for outdoor and indoor acquisitions are 5.6 ms and 400 ms, respectively, highlighting the practical usability of our system. Note that the indoor scene reconstructions (Fig. 6 Scene 3) may appear slightly hazy due to stray light from overhead lighting, as the prototype setup lacks a closed lens barrel. Additionally, a small area of the sensor was stained during the capture process, resulting in a dark patch in some scenes' captures and reconstructions.

6 Conclusion

We introduce a multi-aperture color-coded snapshot hyperspectral imaging system that leverages learned array diffractive lenses, aperture-wise color filters, and a reconstruction network. This multi-aperture design encodes spatial and spectral information independently across color channels, enabling accurate compressive hyperspectral recovery. As a result, we achieve higher fidelity spectral and RGB reconstructions than existing single-diffractive lens and coded-aperture solutions in diverse indoor and outdoor scenes.

Looking ahead, the adaptability of our system to support various array sizes and the simplicity of its underlying design suggest a broad potential for future research. Incorporating non-rotationally symmetric lens representations, as discussed by Liu et al. [2022a], into the design of DOEs could further increase the design space of optical encoding options. Furthermore, incorporating off-axis diffractive propagation, in conjunction with concepts from compound-eye imaging, could enhance wide field-of-view capabilities – unlocking HSI imaging with applications across domains.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61925504, 61621001, 62105243), Major projects of Science and Technology Commission of Shanghai (17JC1400800), Special Development Funds for Major Projects of Shanghai Zhangjiang National Independent Innovation Demonstration Zone (ZJ2021-ZD-008), Fundamental Research Funds for the Central Universities, Shanghai Municipal Science and Technology Major Project (2021SHZDZX0100), NSF Career Award (2047359), Packard Foundation Fellowship, Sloan Research Fellowship, Sony Young Faculty Award, and Research Grants Council of Hong Kong (ECS 27212822, GRF 17208023). The authors thank Xin Liu, Edmund Lam, and He Huang for fruitful discussions.

References

Boaz Arad and Ohad Ben-Shahar. 2016. Sparse Recovery of Hyperspectral Signal from Natural RGB Images. In *European Conference on Computer Vision*. Springer, 19–34.

Boaz Arad, Radu Timofte, Ohad Ben-Shahar, Yi-Tun Lin, and Graham D Finlayson. 2020. Ntire 2020 challenge on spectral reconstruction from an rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 446–447.

Henry Arguello, Samuel Pinilla, Yifan Peng, Hayato Ikoma, Jorge Bacca, and Gordon Wetzstein. 2021. Shift-variant color-coded diffractive spectral imaging system. *Optica* 8, 11 (2021), 1424–1434.

Yitzhak August, Chaim Vachman, Yair Rivenson, and Adrian Stern. 2013. Compressive hyperspectral imaging by random separable projections in both the spatial and the spectral domains. *Appl. Opt.* 52, 10 (Apr 2013), D46–D54.

Seung-Hwan Baek, Hayato Ikoma, Daniel S Jeon, Yuqi Li, Wolfgang Heidrich, Gordon Wetzstein, and Min H Kim. 2021. Single-shot Hyperspectral-Depth Imaging with Learned Diffractive Optics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2651–2660.

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

Andreas Brückner, Jacques Duparré, Robert Leitel, Peter Dannberg, Andreas Bräuer, and Andreas Tünnermann. 2010. Thin wafer-level camera lenses inspired by insect compound eyes. *Optics Express* 18, 24 (2010), 24379–24394.

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

Yuanhao Cai, Jing Lin, Zudi Lin, Haoqian Wang, Yulun Zhang, Hanspeter Pfister, Radu Timofte, and Luc Van Gool. 2022b. MST++: Multi-stage Spectral-wise Transformer for Efficient Spectral Reconstruction. In *CVPRW*.

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

A. Chakrabarti and T. Zickler. 2011. Statistics of Real-World Hyperspectral Images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 193–200.

Praneeth Chakravarthula, Jipeng Sun, Xiao Li, Chenyang Lei, Gene Chou, Mario Bijelic, Johannes Froesch, Arka Majumdar, and Felix Heide. 2023. Thin On-Sensor Nanophotonic Array Cameras. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–18.

Siyu Dong, Hongfei Jiao, Ganghua Bao, Jinlong Zhang, Zhanshan Wang, and Xinbin Cheng. 2018. Origin and compensation of deposition errors in a broadband antireflection coating prepared using quartz crystal monitoring. *Thin Solid Films* 660 (2018), 54–58.

Xiong Dun, Hayato Ikoma, Gordon Wetzstein, Zhanshan Wang, Xinbin Cheng, and Yifan Peng. 2020. Learned rotationally symmetric diffractive achromat for full-spectrum computational imaging. *Optica* 7, 8 (2020), 913–922.

Paul Green, Wenyang Sun, Wojciech Matusik, and Fredo Durand. 2007. Multi-aperture photography. In *Acm Siggraph 2007 Papers*. 68–es.

Peter Hinderberger, Sascha Grusche, and Martin J Losekamm. 2023. Double-dispersive spatio-spectral scanning for hyperspectral Earth observation. *Optica* 10, 6 (2023), 740–751.

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

Hayato Ikoma, Cindy M Nguyen, Christopher A Metzler, Yifan Peng, and Gordon Wetzstein. 2021. Depth from defocus with learned optics for imaging and occlusion-aware depth estimation. In *2021 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 1–12.

Daniel S Jeon, Seung-Hwan Baek, Shinyoung Yi, Qiang Fu, Xiong Dun, Wolfgang Heidrich, and Min H Kim. 2019. Compact snapshot hyperspectral imaging with diffracted rotation. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–13.

Sarawak Kuching. 2007. The performance of maximum likelihood, spectral angle mapper, neural network and decision tree classifiers in hyperspectral image analysis. *Journal of Computer Science* 3, 6 (2007), 419–423.

Anat Levin, Rob Fergus, Frédo Durand, and William T Freeman. 2007. Image and depth from a conventional camera with a coded aperture. *ACM transactions on graphics (TOG)* 26, 3 (2007), 70–es.

Chengbo Li, Ting Sun, Kevin F Kelly, and Yin Zhang. 2012. A compressive sensing and unmixing scheme for hyperspectral data processing. *IEEE Transactions on Image Processing* 21, 3 (2012), 1200–1210.

Lingen Li, Lizhi Wang, Weitao Song, Lei Zhang, Zhiwei Xiong, and Hua Huang. 2022. Quantization-Aware Deep Optics for Diffractive Snapshot Hyperspectral Imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19780–19789.

Yuqi Li, Qiang Fu, and Wolfgang Heidrich. 2021. Multispectral illumination estimation using deep unrolling network. (2021), 1–8.

Chia-Hsiang Lin, Shih-Hsiu Huang, Ting-Hsuan Lin, and Pin Chieh Wu. 2023. Metasurface-empowered snapshot hyperspectral imaging with convex/deep (CODE) small-data learning theory. *Nature communications* 14, 1 (2023), 6979.

Guoxuan Liu, Ning Xu, Huaidong Yang, Qiaofeng Tan, and Guofan Jin. 2022b. Miniaturized structured illumination microscopy with diffractive optics. *Photonics Research* 10, 5 (2022), 1317–1324.

Xin Liu, Linpei Li, Xu Liu, Xiang Hao, and Yifan Peng. 2022a. Investigating deep optics model representation in affecting resolved all-in-focus image quality and depth estimation fidelity. *Optics Express* 30, 20 (2022), 36973–36984.

Yuwei Liu, Hongbin Pu, and Da-Wen Sun. 2017. Hyperspectral imaging technique for evaluating food quality and safety during various processes: A review of recent applications. *Trends in food science & technology* 69 (2017), 25–35.

Bing Lu, Phuong D Dao, Jiangui Liu, Yuhong He, and Jiali Shang. 2020. Recent advances of hyperspectral imaging technology and applications in agriculture. *Remote Sensing* 12, 16 (2020), 2659.

Guolan Lu and Baowei Fei. 2014. Medical hyperspectral imaging: a review. *Journal of biomedical optics* 19, 1 (2014), 010901.

Zacarias Malacara, Daniel Malacara-Hernández, and Zacarías Malacara-Hernández. 2003. *Handbook of optical design*. CRC press.

Ziyi Meng, Zhenming Yu, Kun Xu, and Xin Yuan. 2021. Self-supervised neural networks for spectral snapshot compressive imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2622–2631.

Deniz Mengu, Anika Tabassum, Mona Jarrahi, and Aydogan Ozcan. 2023. Snapshot multispectral imaging using a diffractive optical network. *Light: Science & Applications* 12, 1 (2023), 86.

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

Elias Nehme, Daniel Freedman, Racheli Gordon, Boris Ferdman, Lucien E Weiss, Onit Alalouf, Tal Naor, Reut Orange, Tomer Michaeli, and Yoav Shechtman. 2020. DeepSTORM3D: dense 3D localization microscopy and PSF design by deep learning. *Nature methods* 17, 7 (2020), 734–740.

Yifan Peng, Qiang Fu, Felix Heide, and Wolfgang Heidrich. 2016. The diffractive achromat full spectrum computational imaging with diffractive optics. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–11.

Yifan Peng, Qilin Sun, Xiong Dun, Gordon Wetzstein, Wolfgang Heidrich, and Felix Heide. 2019. Learned large field-of-view imaging with thin-plate optics. *ACM Transactions on Graphics* 38, 6 (2019), 3356526.

Shijie Rao, Yidong Huang, Kaiyu Cui, and Yali Li. 2022. Anti-spoofing face recognition using a metasurface-based snapshot hyperspectral image sensor. *Optica* 9, 11 (Nov 2022), 1253–1259. https://doi.org/10.1364/OPTICA.469653

Gaurav Sharma, Wencheng Wu, and Edul N Dalal. 2005. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur* 30, 1 (2005), 21–30.

Zheng Shi, Yuval Bahat, Seung-Hwan Baek, Qiang Fu, Hadi Amata, Xiao Li, Praneeth Chakravarthula, Wolfgang Heidrich, and Felix Heide. 2022. Seeing through obstructions with diffractive cloaking. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–15.

Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. 2018. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–13.

Rahul Sukthankar, Robert G Stockton, and Matthew D Mullin. 2001. Smarter presentations: Exploiting homography in camera-projector systems. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, Vol. 1. IEEE, 247–253.

Qilin Sun, Congli Wang, Fu Qiang, Dun Xiong, and Heidrich Wolfgang. 2021. End-to-end complex lens design with differentiable ray tracing. *ACM Trans. Graph* 40, 4 (2021), 1–13.

Jun Tanida, Tomoya Kumagai, Kenji Yamada, Shigehiro Miyatake, Kouichi Ishida, Takashi Morimoto, Noriyuki Kondou, Daisuke Miyazaki, and Yoshiki Ichioka. 2001. Thin observation module by bound optics (TOMBO): concept and experimental verification. *Applied optics* 40, 11 (2001), 1806–1813.

Ethan Tseng, Ali Mosleh, Fahim Mannan, Karl St-Arnaud, Avinash Sharma, Yifan Peng, Alexander Braun, Derek Nowrouzezahrai, Jean-Francois Lalonde, and Felix Heide. 2021. Differentiable compound optics and processing pipeline optimization for end-to-end camera design. *ACM Transactions on Graphics (TOG)* 40, 2 (2021), 1–19.

Kartik Venkataraman, Dan Lelescu, Jacques Duparré, Andrew McMahon, Gabriel Molina, Priyam Chatterjee, Robert Mullis, and Shree Nayar. 2013. Picam: An ultra-thin high performance monolithic camera array. *ACM Transactions on Graphics (TOG)* 32, 6 (2013), 1–13.

Congli Wang, Ni Chen, and Wolfgang Heidrich. 2022. dO: A differentiable engine for deep lens design of computational imaging systems. *IEEE Transactions on Computational Imaging* 8 (2022), 905–916.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence* 43, 10 (2020), 3349–3364.

Zhiwei Xiong, Lizhi Wang, Huiqun Li, Dong Liu, and Feng Wu. 2017. Snapshot hyperspectral light field imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3270–3278.

Nan Xu, Hao Xu, Shiqi Chen, Haiquan Hu, Zhihai Xu, Huajun Feng, Qi Li, Tingting Jiang, and Yueting Chen. 2023. Snapshot hyperspectral imaging based on equalization designed DOE. *Optics Express* 31, 12 (2023), 20489–20504.

Motoki Yako, Yoshikazu Yamaoka, Takayuki Kiyohara, Chikai Hosokawa, Akihiro Noda, Klaas Tack, Nick Spooren, Taku Hirasawa, and Atsushi Ishikawa. 2023. Video-rate hyperspectral camera based on a CMOS-compatible random array of Fabry–Pérot filters. *Nature Photonics* 17, 3 (2023), 218–223.

F. Yasuma, T. Mitsunaga, D. Iso, and S.K. Nayar. 2008. *Generalized Assorted Pixel Camera: Post-Capture Control of Resolution, Dynamic Range and Spectrum*. Technical Report.

Bo Zhang, Xin Yuan, Chao Deng, Zhihong Zhang, Jinli Suo, and Qionghai Dai. 2022. End-to-end snapshot compressed super-resolution imaging with deep optics. *Optica* 9, 4 (2022), 451–454.

Jian Zhang, Xiong Dun, Jingyuan Zhu, Zhanyi Zhang, Chao Feng, Zhanshan Wang, Wolfgang Heidrich, and Xinbin Cheng. 2023a. Large Numerical Aperture Metalens with High Modulation Transfer Function. *ACS Photonics* (2023).

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

Wenyi Zhang, Hongya Song, Xin He, Longqian Huang, Xiyue Zhang, Junyan Zheng, Weidong Shen, Xiang Hao, and Xu Liu. 2021. Deeply learned broadband encoding stochastic hyperspectral imaging. *Light: Science & Applications* 10, 1 (2021), 1–7.

Weihang Zhang, Jinli Suo, Kaiming Dong, Lianglong Li, Xin Yuan, Chengquan Pei, and Qionghai Dai. 2023b. Handheld snapshot multi-spectral camera at tens-of-megapixel resolution. *Nature Communications* 14, 1 (2023), 5043.

Yuzhi Zhao, Lai-Man Po, Qiong Yan, Wei Liu, and Tingyu Lin. 2020. Hierarchical regression network for spectral reconstruction from RGB images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 422–423.