DD-Net: spectral imaging from a monochromatic dispersed and diffused snapshot

JONATHAN HAUSER[1†*], AMIT ZELIGMAN[1†], AMIR AVERBUCH[2], VALERY A. ZHELUDEV[2] AND MENACHEM NATHAN[1].

[1]*School of Electrical Engineering, Faculty of Engineering, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel*
[2]*School of Computer Science, Faculty of Exact Sciences, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel*

[†] *These authors contributed equally to this paper.*
[*]*Corresponding author: hauserjo@post.tau.ac.il.*

Abstract: We propose a snapshot spectral imaging (SSI) method for the visible spectral range using a single monochromatic camera equipped with a two-dimensional (2D) binary-encoded phase diffuser placed at the pupil of the imaging lens and by resorting to Deep Learning (DL) algorithms for signal reconstruction. While spectral imaging was shown to be feasible using two cameras equipped with a single, one-dimensional (1D) binary diffuser and Compressed Sensing (CS) algorithms [Appl. Opt. XX, XXX, (XXXX), accepted for publication on July 31st, 2020], the suggested diffuser design expands the optical response and creates optical spatial and spectral encoding along both dimensions of the image sensor. To recover the spatial and spectral information from the dispersed and diffused (DD) monochromatic snapshot, we developed novel DL algorithms, dubbed as DD-Nets, which are tailored to the unique response of the optical system which includes either a 1D or a 2D diffuser. high-quality reconstructions of the spectral cube in simulation and lab experiments are presented for system configurations consisting of a single monochromatic camera with either a 1D or a 2D diffuser. We demonstrate that the suggested system configuration with the 2D diffuser outperforms system configurations with 1D diffuser that utilize either DL-based or CS-based algorithms for the reconstruction of the spectral cube.

1. Introduction

Spectral imaging (SI) and Snapshot Spectral Imaging (SSI) systems have been developed rapidly in recent years thanks to their ability to acquire both spatial and spectral information of an object or a scene. Our previous studies have demonstrated the feasibility of performing SSI and color imaging using a single standard monochromatic digital camera with an added one-dimensional (1D) pupil-domain phase diffuser and a CS-based iterative algorithm [1, 2]. This high-light-throughput architecture is simpler than other computational approaches which used phase-based architectures [3-5] or architectures based on the Coded Aperture Snapshot Spectral Imaging (CASSI) arrangement [6-14]. However, its performance is limited due to discrepancies between the actual optical model and the sensing matrix. This problem was partially mitigated by adding a color RGB camera side-by-side with the monochromatic camera with the diffuser and improving the CS-based algorithm for signal reconstruction [15]. Our more recent study suggested an improved diffuser design which efficiently combines dispersion and diffusion properties using a simple 1D binary phase-encoded profile [16]. While the 1D symmetry of the diffuser enabled complexity relaxation to the reconstruction problem from the dispersed and diffused (DD) image, it introduced only a limited extent of spatial and spectral intermixing of the signal on the sensor. Other issues, which are also common with other CS-based architectures, were sensitivity to mismatches between the mathematical model and the actual physical response of the optical system, long reconstruction times, and sub optimal manual tuning of the algorithm's macro parameters.

Recently, Convolutional Neural Networks (CNNs) were applied as a deep reconstruction algorithm to the field of SI. CNNs are being widely used for various computer vision tasks such image segmentation, classification, and detection [17, 18]. In addition, CNNs are also used for image-to-image reconstruction tasks such as depth reconstruction [19], denoising and inpainting [20], super resolution [21], and image colorization [22]. While CNNs are very commonly used for reconstruction of RGB images, DL methods for hyperspectral images reconstruction are new. An early study for combining a CNN in CS methods used a CNN-based autoencoder architecture to predict a nonlinear representation of the spectral cube to be used as a prior for the spectral image and used it in an iterative

reconstruction process [9]. This method showed high accuracy compared to previous CS techniques [12-13], however it still used iterative optimization methods for reconstruction which lead to long computational times. Wang et al. [10] introduced an optimization-inspired neural network where the spectral prior estimation and the optimization are governed by CNNs. This approach significantly improved the computational time but failed to generalize the reconstruction to real CASSI data. The λ-net, which was recently introduced in [11], presents an end-to-end fully CNN, by using the U-Net framework [23] as a base architecture and combining conditional Generative Adversarial Networks (GANs) [24] and self-attention mechanism [25, 26] for the spectral cube reconstruction from the CASSI measurement. In [27], 3D convolutions were utilized with the U-Net scheme for spectral cube reconstruction from CS measurements.

The introduction of DL algorithms has led to significant improvement in the reconstruction quality, especially on simulated data. However, these approaches still utilize a complex hardware arrangement with reconstruction algorithms that are either prone to relatively long computational times, require manual parameter tuning and yield poor results on real experimental data.

In this paper, we present a reconstruction technique for spectral cube reconstruction of size $256 \times 256 \times 29$ in the $420 - 700nm$ visible spectral range using a single monochromatic camera which is equipped with a (2D) quasi-separable, binary phase diffuser and DL-based algorithms. The diffuser design is a 2D expansion of the 1D binary phase diffuser previously reported by our research group [16].

To reveal additional spatial and spectral information about the scene, the vertical field of view (FOV) is limited, and the dispersion and diffusion of the signal is expanded along both horizontal and vertical dimensions of the sensor. Being also a binary phase element, high light throughput and simple fabrication properties are maintained. To reconstruct the spectral cube solely from the monochromatic DD image, we introduce a novel DL-based algorithm for spectral cube reconstruction, dubbed "DD-Net". Our algorithm leverages the spatial nature of the DD image to fit the NN architecture with a novel, multi-encoder U-Net based architecture followed by variation of self-attention layers crafted to fit the SI reconstruction task. Two variants of the DD-Net algorithm were developed and tested: (i) the "DD1-Net", for a system configuration consisting from a single monochromatic camera with a 1D binary-encoded diffuser [16], and (ii) the "DD2-Net", for a system configuration consisting from a single monochromatic camera with the new, 2D binary-encoded diffuser.

Our simulation and lab experiments indicate that both variants of the DD-Net algorithm achieve high-quality results using a compact hardware arrangement with high optical throughput. While both DD1-Net and DD2-Net algorithms outperform single or dual camera system configurations that use CS-based reconstruction algorithms, we demonstrate that the suggested 2D binary encoded phase diffuser itself has a clear advantage in terms of reconstruction quality with respect to the 1D binary encoded phase diffuser, thanks to the additional information revealed along the vertical dimension of the monochromatic sensor.

2. Mathematical model and arrangement of the optical system with a 2D binary diffuser

The schematic layout of the optical system is shown in Fig. 1.

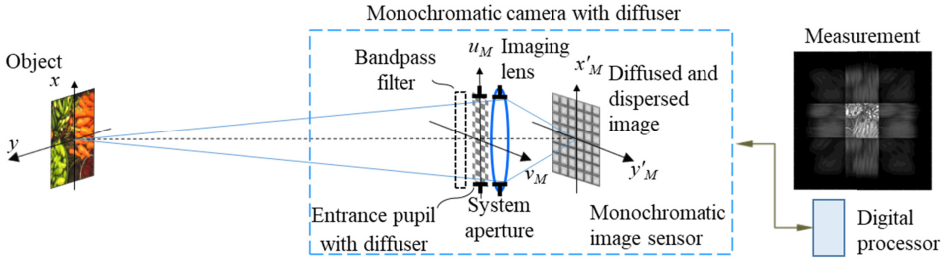

Fig. 1: A Schematic layout of the optical system.

The optical system consists of a monochromatic digital sensor, an imaging lens complete with a wide bandpass filter (BPF), and a transparent, 2D quartz diffuser positioned at the entrance pupil of the imaging lens. This optical arrangement and assembly are similar to the single-camera arrangement described in [1, 2] and to the monochromatic branch of the dual-camera arrangement described in [15, 16], albeit having a 2D binary phase diffuser. The diffuser creates a PSF response that consists of four strong diffusive patterns around the first positive and negative diffraction orders along both horizontal and vertical sensor axes, for each wavelength of interest.

The binary diffuser is designed as a thin phase optical element placed at the entrance pupil of the monochromatic camera with lateral coordinates (u_M, v_M). The diffuser consists of multiple rectangles, in which depth and respective phase are constant within their dimensions and quantized to a binary level. Since the diffuser is placed at the entrance pupil of the imaging lens, it sets a scaled complex pupil function on the exit pupil $P(u'_M, v'_M;\ \lambda_l)$ for each λ_l wavelength of interest:

$$P_M(u'_M, v'_M; \lambda_l) = |P_M(u'_M, v'_M; \lambda_l)| e^{[i\varphi_M(u'_M, v'_M; \lambda_l)]}, \tag{1}$$

where (u'_M, v'_M) are the lateral coordinates of the exit pupil plane, $|P_M(u'_M, v'_M; \lambda_l)|$ is the field amplitude transmittance and $\varphi_M(u'_M, v'_M; \lambda_l)$ is the added phase at the exit pupil. Assuming that the optical system is ideal in the absence of the diffuser, the added phase $\varphi_M(u'_M, v'_M; \lambda_l)$ is solely determined by the diffuser, and the field amplitude transmittance is determined by the clear aperture's shape and size at the exit pupil plane. The incoherent PSF $h_{I,M}$ for the λ_l wavelength can be calculated as

$$h_{I,M}(x'_M, y'_M; \lambda_l) = |FT_{2D}^{-1}\{P_M(u'_M, v'_M; \lambda_l)\}|^2, \tag{2}$$

where $FT_{2D}^{-1}\{\cdot\}$ denotes the inverse 2D Fourier transform and $|\cdot|^2$ denotes the absolute-square operator.

The binary 2D diffuser is intended to provide dispersion and diffusion of light along both horizontal and vertical axes of the image sensor, by setting the PSF of the monochromatic camera. We start by performing 1D binary phase encoding by hard-clipping [16, 28] along the horizontal axis of the diffuser:

$$\varphi_{M,1D}(u'_M; \lambda_{des}) = \begin{cases} \pi, & \text{if } \cos[2\pi f_{enc} \cdot u'_M + \varphi_T(u'_M; \lambda_{des})] > 0 \\ 0, & otherwise \end{cases}, \tag{3}$$

where u'_M is the horizontal coordinate of the exit pupil plane, λ_{des} is the design wavelength, f_{enc} is the encoding spatial frequency that sets the dispersion extent and $\varphi_T(u'_M; \lambda_{des})$ is the encoded phase function that sets the diffusion extent and shape. In Eq. (3) we assume that the diffuser has an ideal field amplitude transmittance [17]. To generate the 2D phase profile of the diffuser, we perform outer product multiplication of the 1D binary phase encoding function along both diffuser axes:

$$\varphi_M(u'_M, v'_M; \lambda_{des}) = \frac{1}{\pi}\varphi_{M,1D}(u'_M; \lambda_{des})\varphi_{M,1D}(v'_M; \lambda_{des}). \tag{4}$$

In our previous publication [16], we examined three types of 1D local diffusion patterns, denoted as Top Hat diffusion, cubic phase diffusion and random phase diffusion. The diffusion patterns were optically dispersed to combine between the requirements of CS theory and classical spectroscopy. Our experimental results indicated that 1D Top Hat diffusion yields best performance on average, for a system configuration that includes a monochromatic camera and an RGB camera and utilizes CS algorithms for spectral cube reconstruction. In this work, we examine a single 2D diffuser that performs local Top Hat diffusion along both horizontal and vertical axes of the monochromatic sensor. Specifically, we set the nominal half-width of the Top Hat profile along each axis to be 22 μm, which equals 10 sensor pixels. The encoding frequency f_{enc} was set to 100 $[mm^{-1}]$, which corresponds to a dispersion extent of ~4.4 pixels per band (4.4 pixels per 10 nm) along both sensor axes. The 2D binary diffuser fabrication was identical to the process described in [16], using standard binary staircase technology on a 2.4 mm thick, 5-inch chrome mask made of quartz.

To calculate the PSF $h_{I,M}(x'_M, y'_M; \lambda_l)$ for each wavelength λ_l of interest, we first calculate the corresponding phase $\varphi_M(u'_M, v'_M; \lambda_l)$ as [1, 16, 29]

$$\varphi_M(u'_M, v'_M; \lambda_l) = \varphi_M(u'_M, v'_M; \lambda_{des})\frac{\lambda_{des}}{\lambda_l}\frac{n(\lambda_l)-1}{n(\lambda_{des})-1}, \tag{5}$$

where $n(\lambda_{des})$ is the refractive index of the diffuser's material for λ_{des}. The PSF is then derived according to Eq. (2). Thus, when added to a regular digital camera, the diffuser converts the original image into a DD image with programmed blur. Fig. 2 shows exemplary analytical calculations of the 2D PSF profiles at 420nm, 550nm and 670nm for the fabricated diffuser, with marked dispersion extent Δ_{disp} and diffusion extent Δ_{diff} at the (1,0) and (0,-1) diffraction orders.

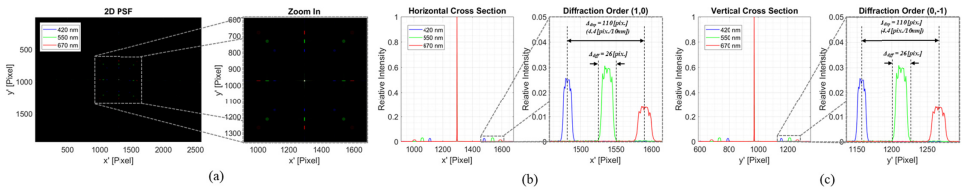

Fig. 2: (a) Experimentally simulated PSF profiles for the encoded Top Hat 2D phase diffuser, for 420nm, 550nm and 670nm; (b) horizontal and (c) vertical cross sections along the middle region of the 2D PSF, with zoom in at the diffusion patterns located in the (1,0) and (0,-1) diffraction orders , with dispersion extent Δ_{disp} and diffusion extent Δ_{diff}.

The intensity $I'_M(x'_M, y'_M;\ \lambda_l)$ of the DD image in the presence of the diffuser at the λ_l wavelength may be expressed by a 2D convolution of the ideal ("non-dispersed") image $I(x, y;\ \lambda_l)$ with the incoherent PSF of the monochromatic camera $h_{I,M}(x'_M;\ \lambda_l)$:

$$I'_M(x'_M, y'_M;\ \lambda_l) = I'_M(\mu_{O,M}x, \mu_{O,M}y;\ \lambda_l) = \iint h_{I,M}(x'_M - x, y'_M - y;\ \lambda_l)\ I(\mu_{O,M}x, \mu_{O,M}y;\ \lambda_l)dxdy, \tag{6}$$

where $\mu_{O,M}$ denotes the optical magnification of the camera. The contribution of polychromatic light to the DD image can be expressed as a weighted sum of intensities of the monochromatic DD images $I'_M(x'_M, y;\ \lambda_l)$ over all wavelengths $\lambda_l, l = \overline{1, L}$, where the weighting coefficients are determined by the spectral sensitivity $S_M(\lambda)$ of the optical system. Hence, the monochromatic DD image signal can be expressed as

$$I'_M(x'_M, y'_M) = \int_{\lambda_a}^{\lambda_b} S_M(\lambda) I'_M(x'_M, y'_M;\ \lambda)\, d\lambda, \tag{7}$$

where $[\lambda_a, \lambda_b]$ is the spectral response range of the optical system. Similar expression can also be derived for the response of the optical system if a 1D diffuser is considered instead of a 2D diffuser [16].

3. Data acquisition for Neural network training

Unlike CS-based algorithms, NN-based algorithms do not require an explicit model of the optical system, but rather a large dataset of measurements acquired from the optical system to perform network training. In the framework of this paper, we explore two methods of data generation for NN training, based on simulation and experimental measurements with our prototype SSI system.

3.1 Simulated Data acquisition

To test our system for reflective scenes and to assess our system performance with respect to other SSI DL-based methods [10, 11, 14, 27, 30, 31], we simulated multiple DD snapshots using the database of spectral cubes from the ICVL HS database [32]. This set includes over 200 spectral cubes of outdoor daily scenes at spatial resolution of 1392×1300 pixels. For each cube, 31 bands were sampled in the spectral domain from 400 to 700 nm with 10nm increment. We used the 420-700nm spectral interval to be consistent with the real data experiment.

For training and validation data, we chose 160 spectral cubes such there is no overlap in the images environment between training/validation set and the test set. We cropped the cubes uniformly into $N_{y,i} \times N_{x,i} \times L = 256 \times 256 \times 29$ patches to create 15000 different cubes. We then simulated the sensor response to generate a DD snapshot measurement for each spectral cube. Specifically, the DD snapshot measurement was calculated according to Eq. (7) for each system configuration with either 1D or the new 2D binary diffusers.

3.2 Real Data acquisition

Experimental acquisition of multiple DD snapshots was done with our prototype with the diffuser, using an Apple iPad Air 2 as an object generator. A collection of 9067 color images was selected from the COCO dataset [33], which contains common objects in their natural context. The images were sequentially projected on the iPad screen, which was connected to a PC unit that synchronized between the projection of an image and the acquisition of the DD snapshot from the prototype. The sensor integration time was set to 7 ms for each object, which guaranteed a non-saturation regime in each DD snapshot. The iPad screen was placed 50 cm away from the prototype and each color image was scaled such that the non-DD image occupied $N_{y,i} \times N_{x,i} = 256 \times 256$ sensor pixels. A single dark measurement was also obtained and later subtracted from all DD snapshots for deducting any ambient signal. To evaluate the performance of the new 2D Top Hat diffuser, we

replaced it with the formerly reported 1D Top Hat diffuser [16] and acquired 9067 DD snapshots of the same objects with the prototype, with an integration time of 4.8 ms. The 1D diffuser was fabricated in the same process as the new 2D diffuser and shared the same spatial encoding (f_{enc} = 100 $[mm^{-1}]$) and nominal half-width of the Top Hat profile (10 sensor pixels), however it introduced the dispersion and diffusion only along the horizontal axis of the sensor. Reference spectral cubes of the 9067 objects were acquired in a separate procedure by the monochromatic camera without a diffuser, using a set of $L = 29$ narrow-bandpass filters at the spectral range of 420-700nm in equal spacing of 10nm, as described in [16]. The integration time of the camera during reference cubes acquisition was 52 ms for each object and for each of the $L = 29$ wavelengths. Fig. 3 summarizes the acquisition schemes for simulation and experimental data.

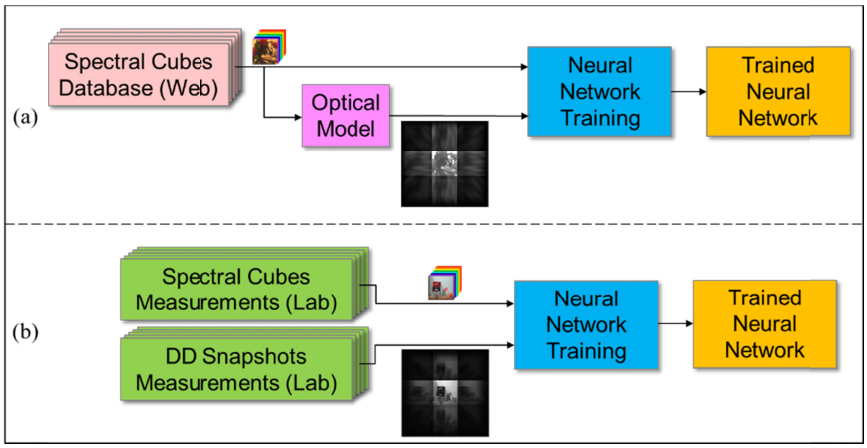

Fig. 3: Simplified block diagrams for (a) simulation and (b) lab experimental datasets acquisition for NN training.

3.3 Data augmentation

The simulated data acquisition process enables to perform data augmentation on the reference spectral cubes and to generate corresponding augmented DD images. This allowed us to perform various data augmentations on the full spatial size (1392×1300) cubes and maintain the same DD image generation process on the augmented cubes. We applied random resize, random rotation, and custom illumination on the full-size cubes, cropped them into $N_{y,i} \times N_{x,i} \times L = 256 \times 256 \times 29$ sub-cubes and calculated the DD images according to Eq. (7). Accordingly, we simulated over 35000 different pairs of spectral cubes and DD snapshots. Augmentation on the real data included solely vertical/horizontal flips on both the spectral cubes and the DD image.

Custom illumination augmentation

To make the predictions agnostic to specific illuminations such as daylight or indoor halogen light, we augmented the natural relations between the spectral channels by simulating a black-body illumination source. Let $y \in \mathbb{R}^{N_{y,i} \times N_{x,i} \times L}$ be a spectral cube with L spectral channels, width $N_{x,i}$ and height $N_{y,i}$. We determine an illumination augmentation as an element-wise product between the spectral cube and an illumination vector, $v \in \mathbb{R}^L$, such that the augmented spectrum $I'_{m,n} \in \mathbb{R}^L$ at the (m,n) spatial pixel is

$$I'_{m,n} = I_{m,n} v \quad , m = 1,2 \ \ldots N_{y,i} \ , n = 1,2 \ \ldots N_{x,i}. \tag{8}$$

The irradiance spectrum of a black body is [34]

$$R(\lambda, T) = \frac{2hc^2}{\lambda^5} \frac{1}{e^{\frac{hc}{\lambda KT}} - 1}, \tag{9}$$

where λ is the radiation wavelength in nanometers, T is the temperature of the black body in Kelvin, h is the Planck constant and c is the light velocity in vacuum.

Simulation of illuminations for a random set of black-body temperatures between 3000-7000K was done according to Eq. (9). For each given black-body temperature, we discretely calculated the radiation vector v in the 420-700nm spectral region of interest and augmented the original spectral cube according to Eq. (8). Fig. 4(a) shows exemplary black-body irradiances used for data augmentation. Fig. 4(b) shows exemplary RGB representations for

spectral cubes which were augmented by black-body illumination sources with temperature of 3000K, 5000K, and 7000K.

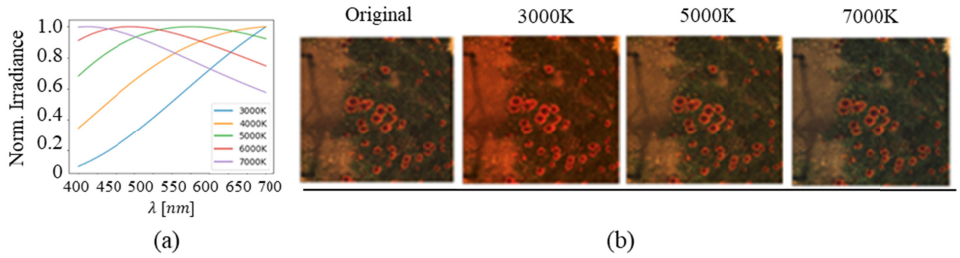

Fig. 4: (a) Normalized radiation vectors at the 400-700nm spectral interval for black-body sources at various temperatures between 3000-7000K; (b) Exemplary RGB representations for spectral cubes which were augmented by black-body illumination sources with temperature of 3000K, 5000K and 7000K.

4. Neural network architectures for spectral cube reconstruction

Reconstruction architecture for a system configuration with a 1D binary diffuser

Data preprocessing

The 1D diffuser creates a DD snapshot which includes a sharp image replica around the zeroth diffraction order (denoted as the "Center Replica") and two strong DD replicas around the -1 and 1 diffraction orders of the horizontal axis (denoted as the "Left Replica" and "Right Replica", respectively). Since the non-DD image is concentric with the optical axis of the camera and has spatial dimensions of 256×256 sensor pixels, the Center Replica also occupies 256×256 pixels around the sensor's center. The Left and Right Replicas occupy a larger footprint due to the dispersion and diffusion introduced by the diffuser. Considering that the optical dispersion is 4.4 pixels per 10 nm, the local diffusion width is 26 pixels, and the spectral range of interest is [420, 700] nm, the dimensions of the Left and Right replicas are 256×405 sensor pixels around the -1 and 1 diffraction order's center of the 560nm wavelength. The Left and Right replicas are located 246 pixels to the left and right to the sensor's center, respectively, and the dimensions of the DD snapshot which includes the 3 replicas are 256×896 pixels. To allow input-output shape correspondence which keeps the fidelity receptive field (the region in the input image that a particular output pixel was calculated from), we cropped the Center Replica of the monochromatic image and concatenated it with the adjacent Left and right Replicas, as illustrated in Fig. 5. Hence, the size of the array which enters the network is $256 \times 256 \times 3$.

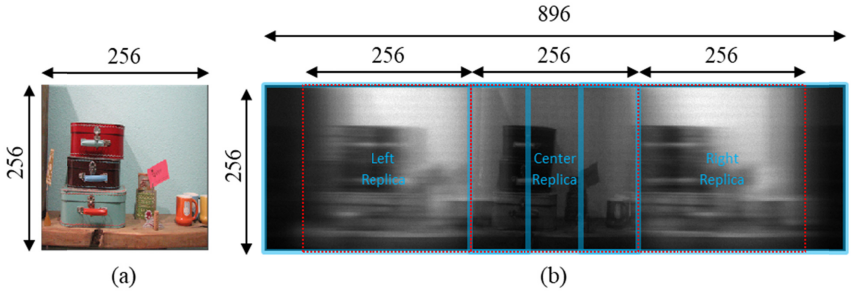

Fig. 5: (a) The image projected on the iPad screen (b) Central region of the DD image which occupies the Left, Center, and Right replicas. The full optical footprints are marked in blue, whereas the cropped regions used for spectral cube reconstruction are marked in dashed-red.

Neural Network architecture

The NN architecture for the reconstruction of the spectral cube from a DD image which was acquired with a 1D diffuser is described in Fig. 6.

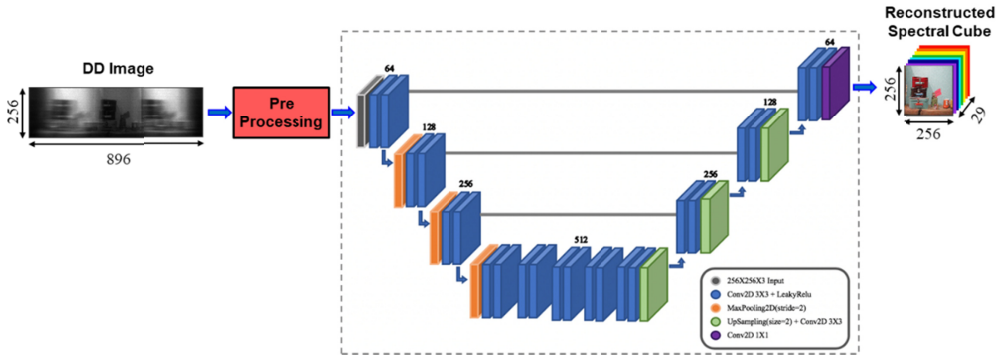

Fig. 6: Network architecture for spectral cube reconstruction from DD images acquired with a system configuration with a 1D diffuser. Number of filters are denoted above the blocks. The gray lines state the skip connections which traverse the features maps from each encoder level into the decoder correspond level. This information is then concatenated with the decoder's previous level to enter the next level of a decoder block.

The skeleton of this architecture is an encoder-decoder scheme in a U-Net structure [23]. The 256x896 DD snapshot is pre-processed to a $256 \times 256 \times 3$ array, which is encoded to a new, low-dimensional domain. Then, the encoded array is decoded to the desired output of size $256 \times 256 \times 29$. The encoder reduces the spatial resolution of the input image along the convolution layers while expanding the number of channels per each feature map. The decoder recovers back the spatial resolution while compressing the number of channels. The skip connections of the U-Net structure preserve the important spatial information in the encoder and traverse it directly to the decoder in the correspond layer. All the NN convolution kernel shapes are 3×3.

Loss function

Common loss functions in image to image reconstruction are either the L2 loss (MSE) or L1 loss (MAE). However, these loss functions do not preserve the natural characteristics of the image and tend to converge to an average value when dealing with multi-dimensional signals. Therefore, we combined 5 different terms in our loss function, which suits the problem of spectral cube reconstruction:

Pixel scaled RMSE Loss: To encourage pixelwise reconstruction fidelity (which is also scale invariant between the reference and reconstructed signals in the objective function) we used a normalized version of the RMSE function:

$$L_{PRMSE}(y,\hat{y}) = \sqrt{\sum_{l}^{L}\sum_{i,j}^{N_{x,i},N_{y,i}}\left(\frac{y_{i,j}^{l}}{\max_{l}\{y_{i,j}\}} - \frac{\hat{y}_{i,j}^{l}}{\max_{l}\{\hat{y}_{i,j}\}}\right)^{2}}, \tag{10}$$

where $\max_{l}\{z_{i,j}\}$ is the maximal value of the spectrum $z_{i,j}$ at the pixel (i,j) over all the $l = 1,..,L$ spectral bands.

SSIM Loss: The Structural Similarity Index Matching (SSIM) [35] loss function $L_{SSIM}(y,\hat{y})$ was utilized on each spectral channel's image to maintain visual similarity.

RGB Loss: The RGB representation of a spectral cube in the visible range provides a precise-spatial and a coarse-spectral evaluation about the scene of interest. Hence, enhanced fidelity between the reference and reconstructed spectral cubes was obtained by utilizing a loss function between their RGB representations [36], y_{RGB} and $\hat{y}_{RGB}$, respectively. The RGB loss $L_{RGB}(y,\hat{y})$ was calculated as the SSIM loss between y_{RGB} and $\hat{y}_{RGB}$:

$$L_{RGB}(y,\hat{y}) = L_{SSIM}(y_{RGB},\hat{y}_{RGB}). \tag{11}$$

Blur Loss: To maintain the spatial sharpness of the reconstructed spectral cube, we utilized a blur loss $L_{blur}(y,\hat{y})$, which is calculated as the absolute difference of the variance of the 2D spatial Laplacian operator on the reference and reconstructed spectral cubes:

$$L_{blur}(y,\hat{y}) = |\,var(\Delta y) - var(\Delta\hat{y})|. \tag{12}$$

Spectral Loss: In order to increase the reconstruction accuracy over the spectral domain, we have utilized a spectral loss term $L_{spectral}(y, \hat{y})$. The spectral loss term is calculated as the absolute difference between the spectral gradient of the reference and reconstructed spectral cubes:

$$L_{spectral}(y, \hat{y}) = \left| \nabla_\lambda y - \nabla_\lambda \hat{y} \right| \tag{13}$$

The final expression for the loss function is:

$$Loss = \lambda_{PRMSE} L_{PRMSE} + \lambda_{RGB} L_{RGB} + \lambda_{SSIM} L_{SSIM} + \lambda_{blur} L_{blur} + \lambda_{spectral} L_{spectral}\ , \tag{14}$$

where the coefficients were set empirically to: $\lambda_{PRMSE} = 0.5,\ \lambda_{RGB} = 0.2,\ \lambda_{SSIM} = 0.1,\ \lambda_{blur} = 0.1$ and $\lambda_{spectral} = 0.1$.

Reconstruction architecture for a system configuration with a 2D binary diffuser

Data preprocessing

Unlike the 1D diffuser, the 2D diffuser is designed to create additional dispersed and diffused image replicas along the horizontal and vertical axes of the sensor. Therefore, the DD image of the 2D diffuser includes additional two strong replicas around the -1 and 1 diffraction orders of the vertical axis (denoted as the "Bottom Replica" and the "Top Replica", respectively), with dimensions of 405×256 sensor pixels around the -1 and 1 diffraction order's center of the 560nm wavelength. Since the design of the 2D diffuser is separable, the Left, Center, and Right replicas of the DD snapshot obtained with a 2D diffuser has the same dimensions and locations as a DD snapshot obtained with a 1D diffuser. The overall dimensions of the DD snapshot which includes 5 replicas of interest are 896×896 pixels. We note that there are 4 additional diagonal replicas which exist around the Center Replica of the DD image. These diagonal replicas are resulted by the "cross terms" of the horizontal and vertical diffraction orders, however they are not used for the reconstruction of the spectral cube due to their weak signal-to-noise level. Fig. 7 shows the DD image obtained with the 2D diffuser, with marked 5 replicas used for the reconstruction of the spectral cube.

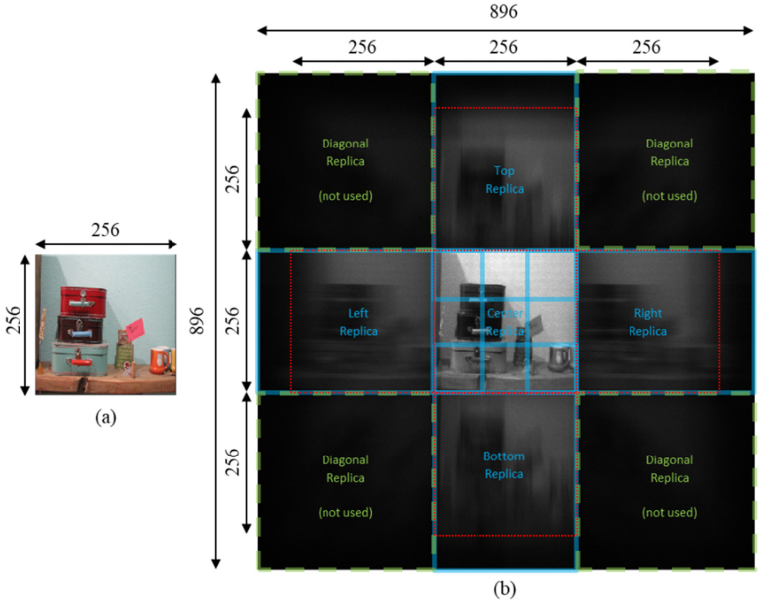

Fig. 7: (a) The image projected on the iPad screen (b) Central region of the DD image which occupies the Left, Center, Right, Top, and Bottom replicas. The full optical footprints are marked in blue, whereas the cropped regions used for spectral cube reconstruction are marked in dashed-red. Diagonal Replicas, marked in dashed green, are not used for signal reconstruction due to their weak signal-to-noise level.

Neural Network architecture

To leverage the additional information provided by the dispersion and diffusion along the vertical axis, we have adapted the NN architecture to extract features along both the horizontal and vertical dimensions. We have therefore developed the multi encoder architecture which utilizes 3 different encoders in parallel which process the DD image as follows:

Horizontal encoder: The inputs of the Horizontal encoder are the Left, Right and Center Replicas, concatenated together into a $256 \times 256 \times 3$ input array. We set the kernel size to be W=9, H=1 to extract only the horizontal spectral features.

Vertical encoder: The Vertical encoder inputs are the Top, Bottom and Center Replicas, concatenated together into a $256 \times 256 \times 3$ input array. The kernel has vertical symmetry (W=1, H=9) to extract the vertical spectral features.

Central encoder: The Central encoder inputs are the Left, Right, Top, Bottom and Center Replicas, concatenated together into a $256 \times 256 \times 5$ input image. The kernel size is W=3, H=3. The central encoder role is to extract spatial features as texture and edges which cannot be discovered by the vertical and horizontal encoders.

Decoder: Following the U-Net decoder architecture we presented above, the input for each decoder block is the concatenation of the previous decoder block output concatenated to the corresponding skip connection output from the encoder. To decode effectively the information extracted from the multi encoder, the decoder input is the concatenation of the previous decoder layer and all 3 encoders outputs of the same layer, as illustrated in Fig. 8. We denote E_h^l, E_c^l, E_v^l and D^l as the Horizontal, Central and Vertical encoder and decoder blocks for l^{th} layer, respectively. We also denote $\theta_{E_h}^l, \theta_{E_c}^l, \theta_{E_v}^l, \theta_D^l$ as the output feature maps from the l^{th} block of the Horizontal, Central and Vertical encoder and decoder blocks, respectively. Following these notations, the outputs of the encoders and decoder are given by:

$$\theta_{E_h}^l = E_h^l\left(\theta_{E_h}^{l-1}\right),\ \theta_{E_c}^l = E_c^l\left(\theta_{E_c}^{l-1}\right),\ \theta_{E_v}^l = E_v^l\left(\theta_{E_v}^{l-1}\right) \tag{15}$$

$$\theta_D^l = D^l\left(concat\left(\theta_{E_h}^l, \theta_{E_c}^l, \theta_{E_v}^l, \theta_D^{l+1}\right)\right) \tag{16}$$

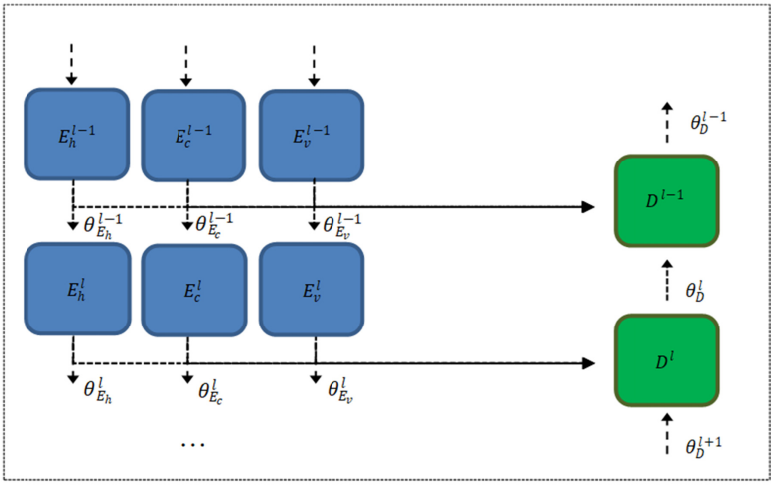

Fig. 8: A block scheme for the $l'th$ layer of the multi encoder-decoder architecture.

Self-attention mechanism

Self-attention, originally aroused in the Natural Language Processing (NLP) field [25], is a mechanism developed to leverage the non-local correlation in the signal to enhance the feature extraction to be globally aware. In vision tasks, where the kernels are very spatially small (usually 3×3), each feature extraction stage 'looks' only on the 3×3 nearest neighbors. Spatial operations such as down-sampling are commonly applied to enlarge the overall receptive field, however the local convolution operations cannot sample a larger region than the dimensions

of the kernel. To overcome this limitation, self-attention layers were adapted for many tasks related to Computer Vision [11, 26].

Considering the suggested optical system with either 1D or 2D binary diffusers, non-local correlation exists between the Center Replica and each of the DD side replicas. While regular convolutions may miss important spectral information encoded in the DD side replicas, the self-attention mechanism enhances the feature extraction from them by being non-locally aware.

Fig. 9 illustrates a self-attention block for a given input features map. In this work, we used not only self-attention blocks, but two more variations of this mechanism, denoted as Vertical Self-Attention and Horizontal Self-Attention blocks. These blocks perform significantly better in both quality and performance with respect to regular self-attention blocks and improve the reconstruction quality of the spectral cube. We first start by describing the self-attention block mechanism.

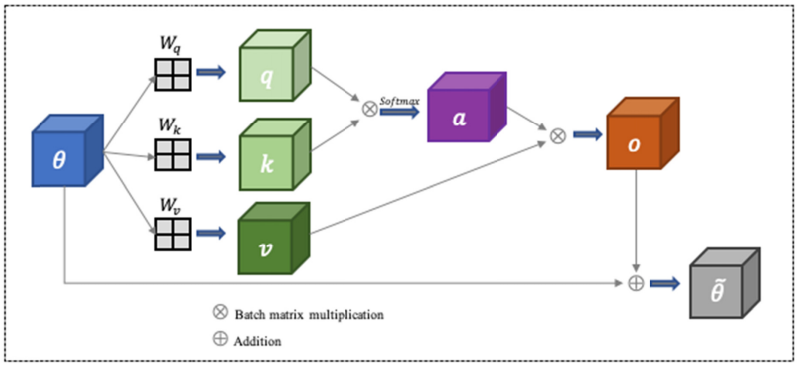

Fig. 9: The self-attention scheme.

Given a features map $\theta \in \mathbb{R}^{H\times W\times C}$ where C is the number of the feature maps and H, W are their height and width respectively, we define the query, key, and value, $q, k, v \in \mathbb{R}^{K\times H\times W\times C'}$ respectively, as:

$$q = W_q * \theta,\ k = W_k * \theta,\ v = W_v * \theta \tag{17}$$

where $W_q,\ W_k, W_v$ are learnable 2D 1×1 convolutions with C' channels and K groups, and $C' = C/K$ where K is the number of attention heads [25]. The output of the self-attention block is a weighted sum of the value tensor where the weights are called attention maps $a \in \mathbb{R}^{K\times H\times W\times H\times W}$ and calculated from the query and key tensors. The element $a_{k,i,j,m,l}$ is representing the weight associated with the k^{th} head of the target pixel $o_{k,i,j,c}$ with respect to the source pixel $v_{k,m,l,c}$ where c is the feature map index and k is the head index. The weighted tensor o is calculated as:

$$o = va^T \in \mathbb{R}^{H\times W\times C},\ o_{k,i,j,c} = \sum_{m=1}^{H}\sum_{l=1}^{W} v_{k,m,l,c}a_{k,i,j,m,l}, \tag{18}$$

where the final dimensions of o are accepted after reshaping the head(K) and channels ($C' = C/K$) dimensions into the original features maps dimensions.

Horizontal and Vertical Self Attention

The regular self-attention approach is trying to exploit the correlation between each of the source pixels to each one of the output pixels. This approach consumes large memory and may be redundant in case most of the pixels are not correlated with each other. However, the presence of the separable diffuser with combination of the weak optical distortion and aberrations of the imaging lens creates strong correlation in the horizontal replicas between pixels in the same row of the DD image, while weak correlation exists among the other pixels. Similarly, the vertical replicas show strong correlation only for pixels on the same column. This optical property led us to develop the Horizontal and Vertical self-attention blocks, where the weights stored in the attention maps are only representing the correlation between a pixel to its corresponding rows or columns.

Consider Horizontal Self-attention block, with attention map $a^h \in \mathbb{R}^{K\times H\times W\times W}$ and target pixel which is correlated only to its self-row pixels. In this setup, we only need to know the column of the source pixel. Therefore, the element $a^h_{k,i,j,l}$ is representing the weight of the target pixel $o^h_{k,i,j,c}$ with respect to the source row $v^h_{k,l,c}$. Similarly,

the element $a^{v}_{k,i,j,m}$ is representing the weight of the target pixel $o^{v}_{k,i,j,c}$ with respect to the source column $v^{v}_{k,m,c}$ in the vertical attention map $a^{v} \in \mathbb{R}^{K\times H\times W\times H}$. Consequently, the target tensors $o^{h}, o^{v} \in \mathbb{R}^{H\times W\times C}$ are calculated as

$$o^{h}_{k,i,j,c} = \sum_{l=1}^{W} v^{h}_{k,l,c} a^{h}_{k,i,j,l}\ ,\quad o^{v}_{k,i,j,c} = \sum_{m=1}^{H} v^{v}_{k,m,c} a^{v}_{k,i,j,m}. \tag{19}$$

All self-attention blocks are utilized right after the Max-Pooling operation in the encoder. Horizontal self-attention blocks were used in the horizontal encoder, and vertical self-attention blocks in the vertical encoder. These blocks leverage the encoded spectral information inside the side replicas of the DD image better than regular self-attention blocks and are much cheaper in terms of memory consumption. To extract spatial information, we also used regular self-attention blocks inside the central encoder. Overall, 9 self-attention blocks from all kinds were used. Fig. 10 illustrates the final architecture of the network with the self-attention blocks.

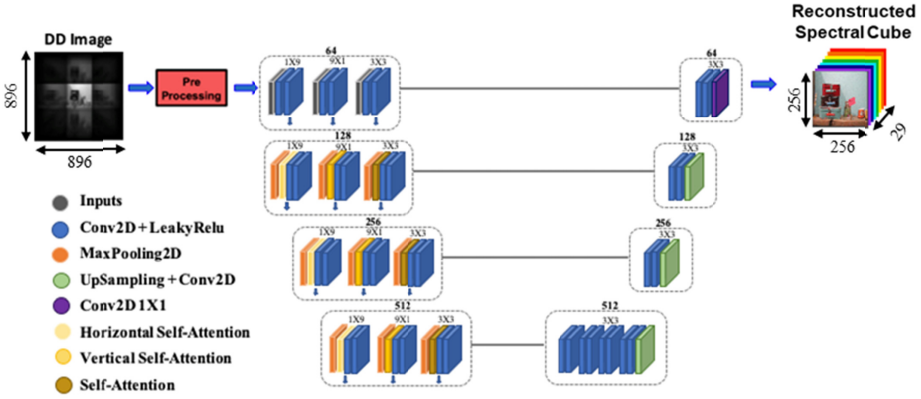

Fig. 10: 2D diffuser reconstruction network with Self-Attention blocks. The number or features map on each block is written in bold. Kernel sizes are written next to each convolution operation. The dashed lines state encoder/decoder blocks. On each of the encoder blocks the horizontal, vertical and central outputs were concatenated together and traverse to the decoder via the skip connections.

5. Results

The training of the NN and its evaluation were performed on a CPU with an Intel Core i-9 processor and 128GB RAM, installed with an Ubuntu OS. The training included 40 epochs with a batch size 8 on a dual Nvidia RTX 2080 TI GPU. We used Adam [37] as an optimizer with a learning rate of 10^{-4}. We reduced the learning rate by a factor of 0.5 every 8 epochs. In all experiments, the evaluation was done on a single Nvidia RTX 2080 TI GPU with a batch size of 1 and the cube dimensions were $256 \times 256 \times 29$. For the DD1-Net, the average reconstruction time per single cube on GPU and CPU was 1.7ms and 0.31s, respectively and for DD2-Net, 6.2ms and 0.64s for GPU and CPU, respectively.

Evaluation on simulated data

The evaluation of our system by simulation was done on 18 test cubes from the ICVL dataset [32]. All test cubes were chosen so their outdoor environment is completely different than the train set. All the 'object' scenes were isolated from the training set, and only part of them were used to evaluate our model. Other outdoor cubes were picked carefully so no similar cubes are within the train set (Consequently, to keep test set fidelity as high as possible, part of the cubes was not used at all). To have a firm comparison of our suggested method with former system configurations, we have performed four types of simulations: (i) a configuration which considered a single monochromatic camera with the 1D binary Top Hat diffuser. A CS-based algorithm, combining Singular Value Decomposition (SVD) and Split-Bregman Iterations (SBI) was used for signal reconstruction, as previously reported by our research group [15,16]; (ii) a configuration which considered a dual camera arrangement (a monochromatic camera and an RGB camera) with the 1D binary Top Hat diffuser placed on the monochromatic camera's lens. The same CS-based, SVD+SBI algorithm, was used for signal reconstruction; (iii) a configuration which again considered a single monochromatic camera with the 1D binary Top Hat diffuser. Here, the new DD1-Net was used for signal reconstruction; (iv) a configuration which considered a single monochromatic camera with the new 2D binary Top Hat diffuser. Here, the new DD2-Net was used for signal reconstruction.

We used 3 metrics, PSNR, SSIM, and the spectral angle mapper (SAM) [38] for quality evaluation. For clarity, we visually display the reconstruction results obtained only by the DD1-Net and the DD2-Net, whereas a comparison summary for all four simulations is given in Table 1. Fig. 11 shows spectral reconstruction for exemplary 3 ICVL scenes. Fig. 11(a) shows the RGB representations of the spectral cubes. Fig. 11(b) shows reference and reconstructed spectra for 3 spatial points of each scene marked in the corresponding row in Fig. 11(a).

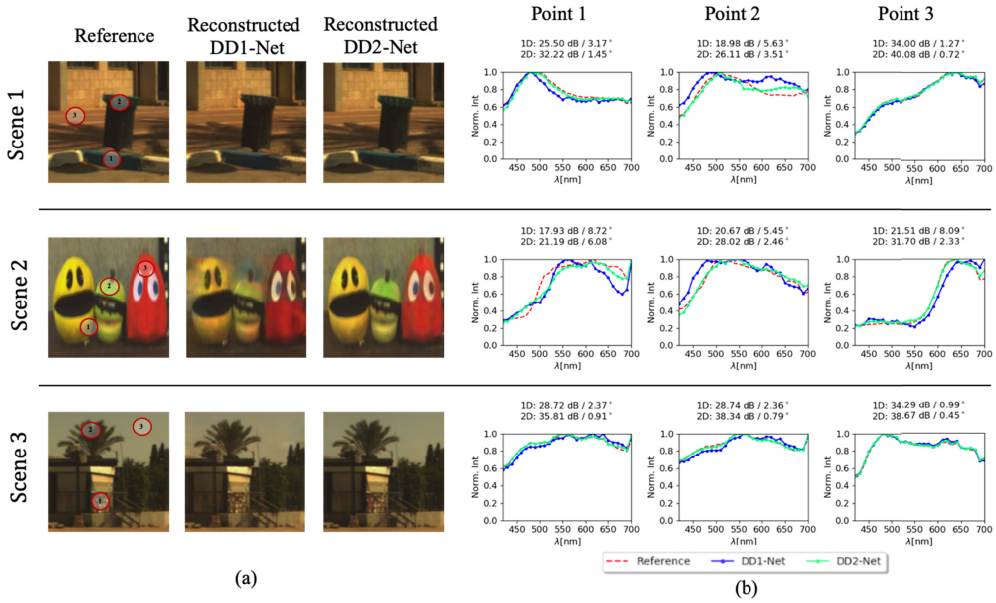

Fig. 11: Simulation results for 3 exemplary scenes using a system configuration with 1D and 2D binary diffusers. (a) RGB representations of the reference and reconstructed spectral cubes obtained by the DD1-Net and DD2-Net algorithms; (b) Reference (dashed Red) and reconstructed spectra for system configurations with 1D binary diffuser and DD1-Net algorithm (blue), and with a 2D binary diffuser and DD1-Net algorithm (light-green) at selected spatial positions as marked in the corresponding reference RGB image to the left, with reconstructions' PSNR and SAM values.

Fig. 12 shows spatial reconstruction for a simulated “Scene 4”. The left column shows the RGB representations of the reference and reconstructed spectral cubes obtained by the DD1-Net and the DD2-Net algorithms, whereas the remaining columns show selected 7 out of 29 reference and reconstructed single-wavelength images, with corresponding PSNR and SSIM values.

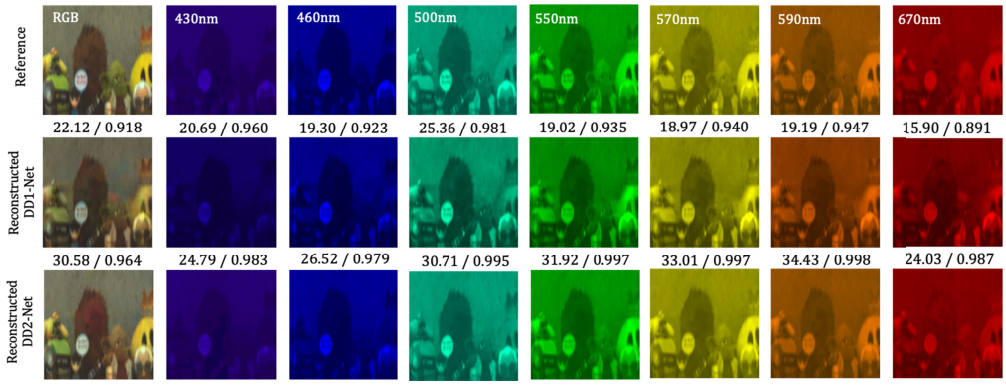

Fig. 12: Simulation results for “Scene 4” using a system configuration with binary 1D and 2D diffusers. Top row: reference images; Middle row: reconstructed images obtained by a system configuration with the 1D binary diffuser and DD1-Net algorithm; Bottom row: reconstructed images obtained by a system configuration with the 2D binary diffuser and DD2-Net algorithm. The left column shows the RGB representations of the spectral cubes, whereas the remaining columns show selected 7 out of 29 reference and reconstructed single-wavelength images, with corresponding PSNR / SSIM values.

The reconstruction results for the simulations with a system configuration with the suggested 2D binary diffuser exhibit excellent performance, as can be indicated from the exemplary scenes shown in Figs. 11 and 12. In

particular, the reconstructed spectra curves trace the corresponding reference spectra curves – and even precisely overlap them. The spatial reconstructions also exhibit remarkable performance, with high PSNR values for most of the spectral channels and nearly ideal SSIM values (~1) along the entire 420-700nm spectral range of interest. Sub-optimal reconstruction performance (e.g., Point 2 in Scene 1, Point 1 in Scene 2, the 430nm, and 670nm single-wavelength images in Scene 4) can be explained with smaller signal-to-noise ratio (SNR) in the reference spectral cubes. Table 1 summarizes the average simulation results on all 18 test cubes.

Table 1. Simulation results summary for all 18 test cubes

System Configuration	PSNR [DB]			SSIM			SAM [DEG]		
	Best	Worst	Mean	Best	Worst	Mean	Best	Worst	Mean
1D diffuser + Mono Camera + SBI&SVD	27.78	20.23	22.71	0.938	0.822	0.874	8.11	12.41	10.02
1D diffuser + Dual Camera + SBI&SVD	27.83	19.50	23.28	0.964	0.910	0.944	7.60	12.72	9.83
1D diffuser + Mono Camera + DD1-Net	37.96	19.62	30.23	0.985	0.887	0.955	1.34	6.35	3.15
2D diffuser + Mono Camera + DD2-Net	**42.07**	**26.51**	**35.56**	**0.993**	**0.956**	**0.980**	**0.97**	**4.30**	**2.16**

Our suggested method, which combines a monochromatic camera with the new 2D binary diffuser and utilizes the DD2-Net reconstruction algorithm, reaches average PSNR of above 35.5 [dB]. This value is significantly higher than the other system configurations which includes the 1D diffuser and utilizes the SVD&SBI or the DD1-Net reconstruction algorithms. The SSIM and SAM metrics further reinforce the reconstruction quality, both in the spatial and the spectral domains. While the system configuration with the 2D diffuser shows a clear advantage over all other system configurations with the 1D diffuser, it can be seen that the DD1-Net outperforms single-camera and even dual-camera system configurations which utilizes the CS-based, SVD&SBI algorithm. This further indicates that our suggested DD-Net architectures are better suited for the task of spectral cube reconstruction from DD snapshots which were obtained in system configurations with either a 1D or a 2D binary diffuser.
We should note that our selected test cubes differ from test cubes used by other CASSI groups for system simulations which used DL-based reconstruction methods dataset. Nevertheless, the average PSNR values obtained by our suggested DD2-Net outperforms the equivalent values obtained by CASSI architectures [10,11,14].

Evaluation on experimental data

Experimental evaluation of our system was done on 17 test cubes acquired from iPad-projected objects. Real data lab experiments were done with the four corresponding system configurations as the simulations, as summarized in Table 2 ahead. Again, we only visually display the reconstruction results obtained by the DD1-Net and the DD2-Net.

Fig. 13 shows spectral reconstruction for each system configuration, for exemplary 3 iPad scenes. Fig. 13(a) shows the RGB representation of the reference spectral cube. Figs. 13(b) and 13(c) the show the RGB representation of the reconstructed spectral cubes for each of the system configurations with 1D and 2D binary diffusers, which were obtained by the DD1-Net and the DD2-Net algorithms, respectively. A red rectangle is marked in each of the RGB images to provide a closer view for selected areas of interest. As can be seen, the system configuration with the 2D diffuser and its corresponding reconstruction architecture outperforms the system configuration with the 1D diffuser, both in terms of spatial and color fidelities. Fig. 13(d) shows reference and reconstructed spectra for 3 spatial points of each scene marked in the corresponding column in Fig. 13(a). The system configuration with the 2D diffuser outperforms the system configuration with the 1D diffuser in terms of spectral reconstruction quality.

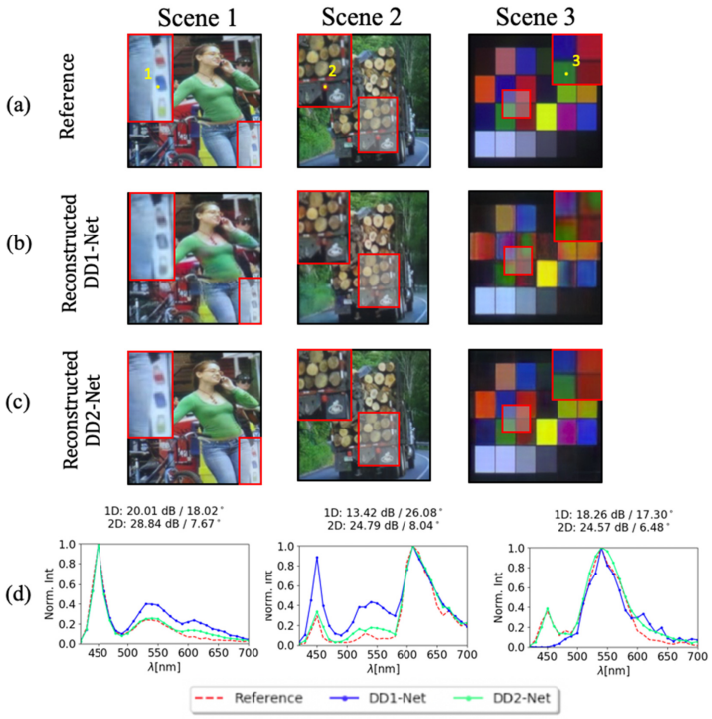

Fig. 13: Lab Experimental results for 3 exemplary scenes using a system configuration with 1D and 2D binary diffusers. (a) RGB representation of the reference spectral cube; (b) and (c) RGB representations of the reconstructed spectral cubes with system configurations with 1D and 2D binary diffusers, obtained by the DD1-Net and the DD2-Net algorithms, respectively. (d) Reference (dashed Red) and reconstructed spectra for system configurations with 1D binary diffuser and DD1-Net algorithm (blue), and with a 2D binary diffuser and DD1-Net algorithm (light-green) at selected spatial positions as marked in the corresponding reference RGB image in Fig 10(a), with reconstructions' PSNR and SAM values.

Fig. 14 shows further spatial reconstruction for experimental "Scene 4". Fig. 14(a) shows the RGB representation of the reference spectral cube, whereas Figs. 14(b) and 14(c) show selected 7 out of 29 reference and reconstructed single-wavelength images, respectively, with corresponding PSNR and SSIM values.

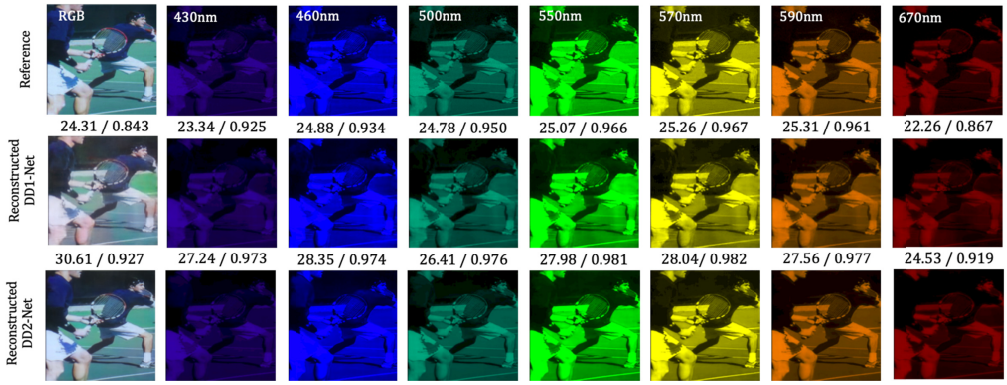

Fig. 14: Experimental results for "Scene 4" using a system configuration with a binary 1D and 2D diffusers. Top row: reference images; Middle row: reconstructed images obtained by a system configuration with the 1D binary diffuser and DD1-Net algorithm; Bottom row: reconstructed images obtained by a system configuration with the 2D binary diffuser and DD2-Net algorithm. The left column shows the RGB representations of the spectral cubes, whereas the remaining columns show selected 7 out of 29 reference and reconstructed single-wavelength images, with corresponding PSNR / SSIM values.

Table 2 summarizes the average experimental results on all 17 test cubes, for each of the system configurations with a 1D or a 2D binary diffuser. The results summary indicates the advantage of the 2D binary diffuser over the 1D binary diffuser, both in terms of spatial and spectral reconstruction quality. The superior quality of the 2D binary diffuser is attributed to the horizontal and vertical directions of dispersion and diffusion, whereas the 1D binary diffuser introduces dispersion and diffusion solely along the horizontal direction of the image sensor. Hence, the 2D diffuser provides richer spatial and spectral information about the scene compared to the 1D diffuser.

The experimental results quality of the 2D diffuser are also attributed to the enhancement of the DD2-Net architecture, which enables to exploit the additional information that is spread along the vertical direction of the sensor. Like our simulations, we also found that the DD1-Net outperforms single-camera and dual-camera system configurations which utilizes the CS-based, SVD&SBI algorithm.

Table 2. Lab Experimental results summary for all 17 test cubes

System Configuration	PSNR [DB]			SSIM			SAM [DEG]		
	Best	Worst	Mean	Best	Worst	Mean	Best	Worst	Mean
1D diffuser + Mono Camera + SBI&SVD	21.60	16.39	19.62	0.566	0.439	0.488	44.24	49.65	46.30
1D diffuser + Dual Camera + SBI&SVD	20.97	13.81	18.25	0.735	0.523	0.614	29.76	44.67	35.70
1D diffuser + Mono Camera + DD1-Net	38.51	30.18	35.22	0.958	0.863	0.930	3.79	13.54	10.07
2D diffuser + Mono Camera + DD2-Net	**42.06**	**35.37**	**38.45**	**0.978**	**0.938**	**0.959**	**3.18**	**12.75**	**7.83**

5. Discussion and conclusions

In this work we presented the design, fabrication, and experiments with a new 2D binary phase diffuser for performing SSI, using a single regular monochromatic camera and by resorting to DL algorithms. The investigated system is an extension of our previously reported SSI system which consisted of a 1D binary diffuser in a dual-camera arrangement [16]. While the former 1D diffuser introduced optical dispersion and diffusion solely along the horizontal axis of the image sensor by utilizing binary phase encoding, the suggested 2D binary diffuser extends the optical response and introduces dispersion and diffusion along both horizontal and vertical axes of the image sensor. Even though this approach limits the field of view along the vertical axis, it provides additional spectral information about the scene which contributes to the reconstruction quality of the spectral cubes, which are sized $256 \times 256 \times 29$ and cover the 420-700 nm visible spectral range. Being consisted of a single monochromatic camera with a binary phase diffuser, our suggested architecture keeps a simple layout with high optical throughput. Consequently, the suggested arrangement is even simpler than our previously reported compact arrangements, consisting of either one camera or two cameras with a 14-level 1D phase diffuser [1, 2, 15], or two cameras with a binary 1D phase diffuser [16].

To recover the spectral cube from the monochromatic DD sensor measurement, we developed novel CNN architectures which are based on the U-Net encoder-decoder scheme, denoted as "DD1-Net" and "DD2-Net". The CNN architectures were adapted for a single snapshot acquired by the monochromatic camera with either a 1D binary diffuser (DD1-Net) or a 2D binary diffuser (DD2-Net) and achieve real-time reconstruction rates. The spatial and spectral information of the scene was reconstructed by processing the sharp Center Replica of the image, with combination of the dispersed and diffused Left and Right Replicas (for either 1D or 2D diffuser) and the Top and Bottom Replicas (for the 2D diffuser only). To leverage the extended distribution of the dispersed and diffused image formed by the 2D binary diffuser, the DD2-Net architecture combines a novel, U-Net based multi-encoder architecture, followed by variation of self-attention layers crafted to fit the SSI reconstruction task. While our system was not directly compared with other groups' computational-SSI architectures such as the popular CASSI architecture [6-14], our simulations on a common database of reflective indoor and outdoor scenes [32] exhibit excellent results using a much simpler hardware arrangement with approximately double optical throughput. Preliminary experimental tests on versatile iPad-projected objects also exhibit very impressive results for a single-camera system arrangement with either 1D or 2D binary diffuser using the DD-Net algorithms. The latter emphasizes the importance of our novel CNN reconstruction architecture, which clearly outperforms CS-based algorithms in terms of reconstruction quality. Moreover, our reconstruction results over 17 test cubes indicate a clear advantage for the system arrangement with the 2D binary diffuser, compared with equivalent reconstructions obtained by either CS-based or DL-based algorithms using a system arrangement with the 1D binary diffuser. The improved performance for the 2D diffuser is explained by the additional spectral and spatial information that is encoded along the vertical dimension and is being used by the neural network for the reconstruction of the spectral cube. Future research may include experiments on real objects, which allow to evaluate more practical scenarios than iPad-projected scenes. Specifically, It should be noted that the spectra of each iPad objects are assembled from a linear combination of three base spectra, which correspond to the unique illumination of the red, green, and blue light emitting diodes (LEDs) of the iPad screen. Since reflectance spectra of real objects greatly differ from the illumination of the iPad's LEDs, we expect that the training phase of the reconstruction network should be done on reflective scenes, rather on iPad objects.

The results of this work indicate further progress towards applications of snapshot spectral imagers in variety of fields, specifically in cases where light throughput, acquisition time, size and cost are cardinal.

6. Funding, acknowledgments, and disclosures

Funding. This research was partially supported by Israeli Ministry of Science Technology and Space [grant numbers 3-10416 and 3-13601].

Disclosures. The authors declare no conflicts of interest.

References

1. M. A. Golub, A. Averbuch, M. Nathan, V. A. Zheludev, J. Hauser, S. Gurevitch, R. Malinsky and A. Kagan, "Compressed sensing snapshot spectral imaging by a regular digital camera with an added optical diffuser", App. Opt. **55** (3), 432-443 (2016).
2. J. Hauser, M. A. Golub, A. Averbuch, M. Nathan, V. A. Zheludev, O Inbar and S. Gurevitch, " High-photon-throughput snapshot colour imaging using a monochromatic digital camera and a pupil-domain diffuser". Journal of Modern Optics 66:7, 710-725, (2019).
3. M. Descour and E. Dereniak, "Computed-tomography imaging spectrometer: experimental calibration and reconstruction results," Appl. Opt., vol. 34, no. 22, pp. 4817–4826, 1995.
4. S. K. Sahoo, D. Tang and C. Dang, 'Single-shot multispectral imaging with a monochromatic camera,' Optica 4, 1209-1213 (2017).
5. P. Wang and R. Menon, "Computational multispectral video imaging," J. Opt. Soc. Am. A 35, 189-199 (2018).
6. M. E. Gehm, R. John, D. J. Brady, R. M. Willett and T. J. Schultz. "Single-shot compressive spectral imaging with a dual-disperser architecture", Opt. Express 15 (21), 14013–14027 (2007).
7. A. Wagadarikar, R. John, R. Willett and D. Brady. "Single disperser design for coded aperture snapshot spectral imaging," Applied Optics 47 (10), B44 – B51 (2008).
8. X. Cao T. Yue, X. Lin, S. Lin, X. Yuan, Q. Dai, L. Carin and D.J Brady. "Computational Snapshot Multispectral Cameras: Toward dynamic capture of the spectral world," in IEEE Signal Processing Magazine, vol. 33, no. 5, pp. 95-108, Sept. 2016. doi: 10.1109/MSP.2016.2582378
9. I. Choi, D. S. Jeon, G. Nam, D. Gutierrez, and M. H. Kim. 2017. High-quality hyperspectral reconstruction using a spectral prior. ACM Trans. Graph. 36, 6, Article 218 (November 2017), 13 pages. DOI:https://doi.org/10.1145/3130800.3130810
10. L. Wang, C. Sun, Y. Fu, M. H. Kim and H. Huang, "Hyperspectral Image Reconstruction Using a Deep Spatial-Spectral Prior," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 8024-8033.
11. X. Miao, X. Yuan, Y. Pu and V. Athitsos, "lambda-Net: Reconstruct Hyperspectral Images From a Snapshot Measurement," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 4058-4068.
12. Yuan, X., Tsai, T. H., Zhu, R., Llull, P., Brady, D., & Carin, L. (2015). Compressive hyperspectral imaging with side information. IEEE Journal of selected topics in Signal Processing, 9(6), 964-976..S.J. Wright, R.D. Nowak, and M.A.T. Figueiredo. 2009. "Sparse reconstruction by separable Approximation". IEEE TSP 57, 7 (2009), 2479–2493
13. Xing Lin, Yebin Liu, Jiamin Wu, and Qionghai Dai. 2014. "Spatial-spectral Encoded Compressive Hyperspectral Imaging". ACM Trans. Graph. (TOG) 33, 6 (2014).
14. Z. Xiong, Z. Shi, H. Li, L. Wang, D. L, F. W, HSCNN: "CNN-Based Hyperspectral Image Recovery from Spectrally Undersampled Projections". ICCV 2019.
15. J. Hauser, M. A. Golub, A. Averbuch, M. Nathan, V. A. Zheludev and M. Kagan, "Dual-camera snapshot spectral imaging with pupil-domain optical diffuser and compressed sensing algorithms". App. Opt. 59 (4), 1058-1070 (2020).
16. J. Hauser, A. Averbuch, M. Nathan, V. A. Zheludev, M. Kagan and M. A. Golub, "Design of binary phase diffusers for a Compressed Sensing snapshot spectral imaging system with two cameras". App. Opt. XX (X), XXX-XXX, accepted for publication on July 31st, 2020.
17. K. He, G. Gkioxari, P. Dollar, R. Girshick, "MaskR-CNN". arXiv:1703.06870, 2018.
18. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Communications of the ACM, vol. 60, no. 6, pp. 84–90, 2017.
19. H. Haim, S. Elmalem, R. Giryes, A. M. Bronstein and E. Marom, "Depth Estimation From a Single Image Using Deep Learned Phase Coded Mask," in *IEEE Transactions on Computational Imaging*, vol. 4, no. 3, pp. 298-310, Sept. 2018, doi: 10.1109/TCI.2018.2849326..
20. J. Xie, L. Xu, E. Chen, "Image Denoising and Inpainting with Deep Neural Networks", NIPS 2012.
21. C. Dong, C. C. Loy, K. He, and X. Tang, "Image Super-Resolution Using Deep Convolutional Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 2, pp. 295–307, 2016.
22. A. K. p, P. R, and M. Ramalaingam, "Image Colorization Using Convolutional Neural Networks," SSRN Electronic Journal, 2019.
23. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," Lecture Notes in Computer Science Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp. 234–241, 2015.
24. P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
25. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In Advances in Neural Information Processing Systems, 2017, pages 6000–6010.
26. Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. "Non-local neural networks". In CVPR, 2018.
27. D. Gedalin, Y. Oiknine, and A. Stern, "DeepCubeNet: reconstruction of spectrally compressive sensed hyperspectral images with deep neural networks," Opt. Express 27, 35811-35822 (2019)
28. S. Shwartz, M. A. Golub and S. Ruschin, "Computer-generated holograms for fiber optical communication with spatial-division multiplexing," Appl. Opt. 56, A31-A40 (2017).
29. M. A. Golub, "Generalized conversion from the phase function to the blazed surface-relief profile of diffractive optical elements," J. Opt Soc. Am. A 116, 1194–1201 (1999).
30. Q. Xie, M. Zhou, Q. Zhao, D. Meng, W. Zuo, Z. Xu, "Multispectral and Hyperspectral Image Fusion by MS/HS Fusion Net". CVPR 2019.

31. Y. Fu, T. Zhang, Y. Zheng, D. Zhang, H. Huang, “Hyperspectral Image Super-Resolution with Optimized RGB Guidance”. CVPR 2019
32. B. Arad and O. Ben-Shahar, Sparse Recovery of Hyperspectral Signal from Natural RGB Images, in the European Conference on Computer Vision, Amsterdam, The Netherlands, October 11–14, 2016.
33. T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: Common Objects in Context. In Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693 (2014).
34. M. Planck (annotated by Hans Kangro; transl. D. ter Haar and Stephen G. Brush): Planck's Original Papers in Quantum Physics (Taylor & Francis, London, 1972; German and English Edition).
35. Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image Quality Assessment: From Error Visibility to Structural Similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
36. D. H. Foster. URL: https://personalpages.manchester.ac.uk/staff/david.foster/Tutorial_HSI2RGB/Tutorial_HSI2RGB.html
37. D.P. Kingma, J.L. Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).
38. F. A. Kruse, A. B. Lefkoff, J. B. Boardman, K. B. Heidebrecht, A. T. Shapiro, P. J. Barloon, and A. F. H. Goetz. "The Spectral Image Processing System (SIPS) - Interactive Visualization and Analysis of Imaging spectrometer Data." Remote Sensing of Environment 44 (1993): 145-163.