

# Compressive spectral video by dynamic spatial-spectral-temporal windowed codification

DAVID MORALES-NORATO<sup>1</sup>, MIGUEL MARQUEZ<sup>1</sup>, ANDRÉS JEREZ<sup>2</sup>,  
HENRY ARGUELLO<sup>3</sup>, ESTEBAN VERA<sup>4</sup>, PABLO MEZA<sup>5,\*</sup>

<sup>1</sup> Department of Physics, Universidad Industrial de Santander, Bucaramanga, Colombia

<sup>2</sup> Department of Electrical Engineering, Universidad Industrial de Santander, Bucaramanga, Colombia

<sup>3</sup> Department of Computer Science, Universidad Industrial de Santander, Bucaramanga, Colombia

<sup>4</sup> Department of Electrical Engineering, Pontificia Universidad Catolica de Valparaiso, Valparaiso, Chile

<sup>5</sup> Department of Electrical Engineering, Universidad de La Frontera, Temuco, Chile

\*[pablo.meza@ufrontera.cl](mailto:pablo.meza@ufrontera.cl)

**Abstract:** Snapshot compressive imaging of high-dimensional data is indispensable for numerous areas of study. Despite the remarkable advances in either spectral or temporal approaches separately, optically compressing a spectral video in a single snapshot remains challenging. In this article, we report a novel compressive spectral video scheme that offers a dynamic color-coded wavelength coding plus a windowing approach to handle the temporal dimension. The proposed compressive coding scheme is implemented by synchronizing a tunable bandpass filter and a coded aperture device to code and acquire the spectral video within a fixed exposure time. Special attention is given to the design and evaluation of the dynamic coded apertures using binary masks. We develop and implement a plug-and-play alternating-direction method of multiplier to efficiently recover the four-dimensional datacubes. We demonstrate the ability of the proposed approach to capture and reconstruct spectral video information in both spectral classification and motion tracking applications.

© 2023 Optica Publishing Group

## 1. Introduction

Spectral video imaging (SVI) aims to record a four-dimensional (4D) data cube  $(x, y, \lambda, t)$ , which contains the intensity information about each spatial-temporal location  $(x, y, t)$  at each wavelength  $\lambda$ . SVI contains information about the time evolution of morphological and spectral features of the scenes that, combined with the mathematical model-based algorithms, sparked the development of high-level applications, e.g., detection and classification applications [1–8]. However, the capture of high-resolution SVI demands high acquisition times, storage, and data transfer rate requirements, which increase as the 4D data cube resolution grows, limiting its widespread usage.

Compressive spectral video imaging (CSV) relies on using compressive sensing theory [9] to drive novel optical designs able to acquire massive high-dimensional datacubes in a single snapshot. In particular, CSV arises as an alternative to acquiring SVIs in high-dynamic scenarios without relying on high-storage capacities and scan-based sensing protocols. A CSV pathfinder work is the coded aperture snapshot spectral imager (CASSI) [10], which aims to reduce the acquisition complexity by optically encoding the SVI's in spatial-spectral dimensions. The CASSI architecture performs a spatial encoding with a spectral sweep in a frame-by-frame fashion, without temporal compression. CASSI-based CSV approaches can be classified into three main categories: hardware design [11, 12], encoding optimization [13], and computational algorithms [11, 12, 14, 15]. From the hardware side, few approaches have focused on extending the compression to include the temporal dimension, often focusing on extending the functionality of compressive spectral imagers to record a compressive measurement per frame. To improve the spatial-temporal reconstruction accuracy, the optical system in [11] incorporates a beam

splitter for dividing and relaying the scene's wavefront into a CASSI system (spectral dimension) and a high-frame-rate panchromatic camera (temporal dimension). Nonetheless, these side-information-based systems require further efforts related to the optical setup calibration process and the development of reconstruction/fusion algorithms. In [12], the optical system uses a LED-based active illumination source for spectral modulation; however, active illumination strategies can lead to bulkier imaging systems, besides presenting calibration robustness issues. The second approach focuses on optimizing the CSVI's encoding optical elements to guarantee that the sensing matrix satisfies one or several inverse properties, e.g., the restricted isometry property (RIP) [16], the Gershgorin circles [17], or the conditional numbers [18]. An optical optimization method for designing colored-coded apertures (CCA) is based on a relaxation of the RIP metric [13], which results in CCAs with uniform sensing in the four dimensions. On the other hand, computational CSVI algorithms were proposed in [14, 15], exploiting the convolutional sparse coding theory. Recently, computational methods based on deep learning (DL) have gained high popularity in spectral imaging reconstruction tasks. The synergistic combination of DL with spectral optical architectures allowed the computational imaging community to introduce new optical systems by designing the underlying coded elements in a data-driven approach [19–22]. Specifically, the two main drawbacks of the DL-based CSVI approaches are the inaccessibility to public SV databases and the elevated computational resources required for training. Despite the progress in multidimensional compression [23], most of the latest research has focused on CSVI optical architectures that avoid performing temporal compression to circumvent the high computational complexity and often poor reconstruction results.

To overcome these limitations, we propose a new snapshot compressive spectral video codification approach. It provides spectral and temporal compression using a dynamic spectral and temporal windowing encoding method, synchronizing a tunable spectral filter with a DMD to achieve spectral dynamic CCAs. Experiments consider a coded aperture design based on a temporal windowed approach to improve the reconstruction performance. For CSV reconstruction, we develop a novel hierarchical-based methodology based on the plug-and-play alternating direction method of multipliers (PnP-ADMM) approach. The proposed method is tested on spectral classification and object detection problems, demonstrating its ability to compressively capture and recover key information from spectral video scenes without requiring capturing the entire datacube beforehand.

## 2. Methodology

### 2.1. Continuous sensing model

The proposed spectro-temporal modulation strategy can be implemented with a liquid crystal tunable bandpass filter (LCTF), followed by a DMD, located in the focal image plane, sequentially accommodated and separated by a relay lens, as can be seen in Fig. 1. The first step is to define the propagation model in continuous notation to develop the discrete version and reconstruction models.

Let  $f_0(x, y, \lambda, t)$  be the spatial-spectral-temporal object, where  $(x, y)$  indexes the spatial coordinates,  $\lambda$  indexes the wavelength, and  $t$  indexes the temporal dimension. An objective lens transmits the source density to the input plane of a  $4f$  system composed of two relay lenses with an LCTF to spectral filter the incident light. In the  $4f$  system's output plane, a DMD synced with the LCTF is located to dynamically modulate each spectral datacube following a 3D color-coded aperture strategy [24]. The resulting wavefront after the first DMD can be expressed as

$$f_1(u, v, \lambda, t) = \kappa_1(u, v, \lambda, t) \iint \gamma(\lambda) f_0(x, y, \lambda, t) h_1(x - u, y - v, t) dx dy, \quad (1)$$

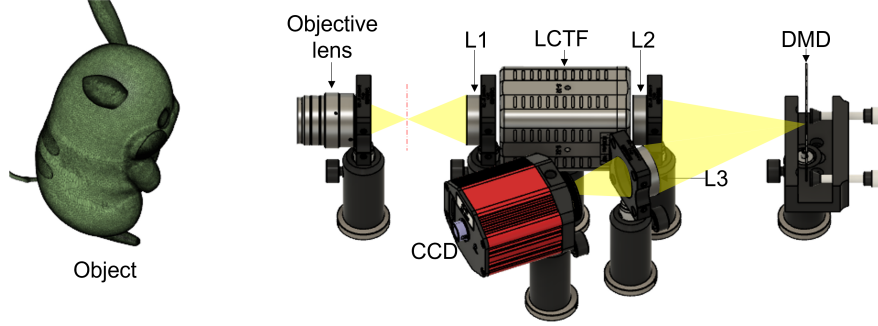


Fig. 1. Sketch of the optical system, which is composed of two spectral and temporal modulation elements LCTF and DMD. Here, the LCTF and the DMD are synced, i.e., they share the same switching speed.

with

$$\kappa(u, v, \lambda, t) = \sum_{i_u, i_v, i_\lambda, i_t} C_{i_u, i_v, i_\lambda, i_t} \cdot \text{rect}\left(\frac{u}{\Delta_c} - i_u, \frac{v}{\Delta_c} - i_v, \frac{\lambda}{\Delta_\lambda} - i_\lambda, \frac{t}{\Delta_t} - i_t\right), \quad (2)$$

where  $h_1(\cdot)$  is the intensity point-spread-function (PSF) introduced by the 4f system with the LCTF located in the middle,  $\text{rect}(\cdot)$  represents a rectangle function,  $\gamma(\cdot)$  represents the spectral source density,  $\mathbf{C} \in \mathbb{R}^{N_x \times N_y \times N_\lambda \times N_t}$  is the binary coding applied to the  $(i_u, i_v, i_\lambda, i_t)$ -th voxel with  $C_{i_u, i_v, i_\lambda, i_t} \in \{0, 1\}$ , and  $\kappa(u, v, \lambda, t)$  represents the continuous version of the mask pattern [referred to as DMD in Fig. 1]. Here,  $i_x \in \{0, \dots, N_x - 1\}$ ,  $i_y \in \{0, \dots, N_y - 1\}$ ,  $i_\lambda \in \{0, \dots, N_\lambda - 1\}$ , and  $i_t \in \{0, \dots, N_t - 1\}$ , where  $N_u$  and  $N_v$  represent the spatial resolution of the mask,  $N_\lambda$  represents the LCTF's total of reachable spectral steps and  $N_t$  is the number of frames. Finally,  $\Delta_c$  represents the coded aperture pixel size,  $\Delta_\lambda$  is the LCTF's spectral resolution, and  $\Delta_t$  represents the temporal resolution related to the switching speed of the LCTF and the DMD elements. The resulting filtered and coded wavefront is propagated and focused into a detector array. The wavefront field immediately before the detector can be expressed as

$$f_2(x', y', \lambda, t) = \iint f_1(u', v', \lambda, t) h_2(u' - x', v' - y') du' dv', \quad (3)$$

where  $h_2(\cdot)$  represents the intensity PSF for the L3 lens. Finally, the measurement at the  $(i_{x'}, i_{y'})$ -th pixel is represented by

$$G_{i_{x'}, i_{y'}} = \int_{\Gamma} \int_{\Lambda} \rho(\lambda) \iint f_2(x', y', \lambda, t) \cdot \text{rect}\left(\frac{x'}{\Delta_d} - i_{x'}, \frac{y'}{\Delta_d} - i_{y'}\right) dx' dy' d\lambda dt, \quad (4)$$

where  $\Lambda$  represents the wavelength axis over the spectral range,  $\Gamma$  represents the time axis over the temporal range,  $\rho(\lambda)$  represents the system's normalized quantum efficiency,  $\Delta_d$  is the size of the camera pixel.

## 2.2. Discrete sensing model

The proposed spectrally dynamic and spatial-temporal windowed codification approach can be schematically represented by Fig. 2.

Imaged by a front optics onto the image plane of a 4f system with a bandpass filter in the middle, the data acquisition starts by filtering the spectral dynamic scene denoted by  $\mathbf{F} \in \mathbb{R}^{N_x \times N_y \times N_\lambda \times N_t}$ , where  $N_x$  and  $N_y$  represent the data lengths in the two spatial dimensions,  $N_\lambda$  and  $N_t$  represents

the data length in the spectral and temporal dimension, respectively. Then, the filtered dynamic scene is imaged into a binary reflective DMD, where the reflected light is spatially encoded (denoted by  $\mathbf{C} \in \mathbb{R}^{N_x \times N_y \times N_\lambda \times N_t}$ ). Note the DMD's pattern changes with each change of the tunable filter, which repeats its cycle  $N_t$ -times. Finally, the resulting spectral and filtered spectral dynamic scene is integrated by the sensor, producing the compressively recorded 2D snapshot

$$\mathbf{G} = \left[ \sum_{i_t=0}^{N_t-1} \sum_{i_\lambda=0}^{N_\lambda-1} \mathbf{F}_{:::,i_\lambda,i_t} \odot \mathbf{C}_{:::,i_\lambda,i_t} \right] + \mathbf{E}, \quad (5)$$

where  $\mathbf{G} \in \mathbb{R}^{N_x \times N_y}$  is the compressed measurements,  $\odot$  represents the element-wise product operation, and  $\mathbf{E} \in \mathbb{R}^{N_x \times N_y}$  represents the additive noise.

### 2.3. Temporal windowing approach

The mask pattern ( $\mathbf{C}$ ), a crucial element in SCI approaches, is designed using a pixel-wise Bernoulli random variable with a transmittance of  $t_c \approx 0.5$ . The spectral depth of the proposed codification approach is limited by the bandpass filter wavelengths; the temporal resolution [i.e.,  $N_\lambda$ ]. The proposed system yields a compression ratio of  $N_\lambda N_t$ . Several works have demonstrated that the compression ratio is highly related to the reconstruction accuracy of CSVI systems [25, 26]. Some works have studied the mask pattern design to improve the reconstruction quality without sacrificing the systems' compression power. However, these designs are limited to three-dimensional scenarios, e.g., compressive spectral or temporal imaging. Inspired by the macro pixel structure approaches and spectral filter arrays, we propose a temporal-macro pixel structure (named windowing strategy) that allows an optimal distribution of the information across the sensor by avoiding large clusters of temporal and spectral information per pixel. In particular, this structure decreases the number of information encoded into each pixel at the expense of the spatial resolution. Introducing the temporal windowing encoding strategy in Eq. (5), we obtain the following model

$$\mathbf{G}_{i_w} = \left[ \sum_{i_t=0}^{S-1} \sum_{i_\lambda=0}^{N_\lambda-1} \mathcal{P}_{i_w} (\mathbf{F}_{:::,i_\lambda,(i_t+i_w S)} \odot \mathbf{C}_{:::,i_\lambda,i_t}) \right] + \mathbf{E}_{i_w}, \quad (6)$$

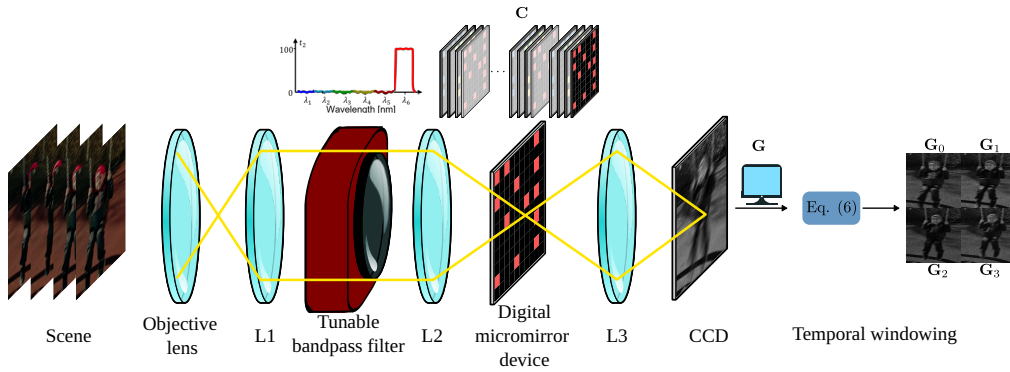


Fig. 2. Illustration of the proposed optical coding methodology with a window size of  $W = 2$ , number of frames  $N_t = 16$ , and temporal interval per pixel of  $S = 4$ . The optical system performs the modulation of the coded aperture  $\mathbf{C}$  by the syncing of the LCTF and DMD. Post-acquisition, the temporal windowing operator  $\mathcal{P}_{i_w}$  separates the data into corresponding  $\mathbf{G}_{i_w}$  measurements.



where  $\mathcal{P}_{i_w}(\cdot) : \mathbb{R}^{N_x \times N_y} \rightarrow \mathbb{R}^{\frac{N_x}{W} \times \frac{N_y}{W}}$  is a subsampling operator [27] with  $i_w \in \{0, \dots, W^2 - 1\}$ , each  $i_w$  measurement contains the information of a temporal interval of size  $S = N_t/W^2 \in \mathbb{N}$ , with  $W \in \mathbb{N}$  are the windowing size, the total of information per pixel in the compressed measurement is relaxed from  $\left(1 - \frac{W^2}{N_\lambda N_t}\right)$  to  $\left(1 - \frac{1}{N_\lambda N_t}\right)$ . Following the temporal macro pixel structure, Eq. (6) can be decoupled into  $W^2$  linear sensing subproblems as

$$\mathbf{g}_{i_w} = \mathbf{H}_{i_w} \mathbf{f}_{i_w} + \boldsymbol{\epsilon}_{i_w}, \quad (7)$$

where  $\mathbf{g}_{i_w} \in \mathbb{R}^m$  represents the compressed measurements related to the temporal interval  $(S \cdot i_w)$ -th frame to  $S \cdot (i_w + 1)$ -th frame with  $m = \frac{N_x N_y}{W^2}$ ;  $\mathbf{f}_{i_w} \in \mathbb{R}^n$  represents the vectorial version of the  $i_w$ -th temporal fragment with  $n = N_x N_y N_\lambda S$ ;  $\boldsymbol{\epsilon}_{i_w} \in \mathbb{R}^m$  is the additive noise for the  $i_w$ -th measurement. The structure of  $\mathbf{H}_{i_w} \in \mathbb{R}^{m \times n}$  relates to the accommodation of the coding elements of the system. In particular, its entries are given by

$$(H_{i_w})_{i,j} = \begin{cases} (c_{i_w})_j, & \text{if } i = \text{mod}(j, m) \\ 0, & \text{otherwise} \end{cases}, \quad (8)$$

for  $i \in \{0, \dots, m-1\}$  and  $j \in \{0, \dots, n-1\}$ , where  $(c_{i_w})_j$  are the entries of  $\mathbf{c}_{i_w} \in \mathbb{R}^n$ , the vectorized version of  $\mathbf{C}_{i_w} = \mathcal{P}_{i_w}(\mathbf{C})$ , and  $\text{mod}(\cdot) : \mathbb{R} \rightarrow \mathbb{W}$  represents the modulus operator.

Additionally, as was explored previously in the state-of-the-art [28–30], to reconstruct high dimensional data, the reconstruction of multiple dimensions could be relaxed by dividing it into substeps. Here, we first reconstruct a grayscale temporal video of the scene, subsequently, the grayscale estimation is used to reconstruct the SV. To achieve this, the proposed sensing methodology defines the coded aperture  $\mathbf{C}$  by merging a 4D random mask  $\mathbf{A} \in \mathbb{R}^{N_x \times N_y \times N_\lambda \times N_t}$  with a 3D random mask  $\mathbf{B} \in \mathbb{R}^{N_x \times N_y \times N_t}$  to encode the temporal domain using an independent mask, which can be represented mathematically as  $(\mathbf{C}_{i_w})_{:::,i_\lambda,i_t} = \mathbf{A}_{:::,i_\lambda,i_t} \odot \mathbf{B}_{:::,i_t}$ . The reconstruction performance depends on the transmittance  $t_c$  of  $\mathbf{C}$  defined as,

$$t_c = \frac{1}{n} \sum_{i_t=0}^{N_t-1} \sum_{i_\lambda=0}^{N_\lambda-1} \sum_{i_x=0}^{N_x-1} \sum_{i_y=0}^{N_y-1} \mathbf{C}_{i_x i_y i_\lambda i_t}. \quad (9)$$

Note that  $t_c = t_a \cdot t_b$ , where  $t_a \in [0, 1]$  and  $t_b \in [0, 1]$  correspond to the transmittance of  $\mathbf{A}$  and  $\mathbf{B}$ , respectively. The optimal values of  $t_a$  and  $t_b$  could be explored to improve the reconstruction performance. Figure 3 illustrates an example of this coding structure in the spectral-temporal dimension composed by the spectral response of the bandpass filter and the codification terms for any spatial pixel.

Imposing the property of  $\frac{1}{N_\lambda} \sum_{i_\lambda=0}^{N_\lambda-1} \mathbf{A}_{:::,i_\lambda,:} = \mathbf{1}$ , and the measurement  $\mathbf{g}_{i_w}$  from Eq. (7) is possible to formulate a inverse problem to a grayscale version  $\mathbf{z}_{i_w}$  of the SV  $\mathbf{f}_{i_w}$  as,

$$\mathbf{g}_{i_w} = \mathbf{H}_{i_w} \mathbf{D}^T \mathbf{z}_{i_w}, \quad (10)$$

where  $\mathbf{z}_{i_w} \in \mathbb{R}^{N_x N_y S}$  is a grayscale video representation of the  $i_w$ -th spectral-video segment with  $\mathbf{z}_{i_w} = \mathbf{D} \mathbf{f}_{i_w}$  and  $\mathbf{D} = \mathbf{I}_{S \times S} \otimes \left[ \mathbf{1}_{N_\lambda}^T \otimes \mathbf{I}_{\frac{N_x N_y}{W^2} \times \frac{N_x N_y}{W^2}} \right]$  is a spectral downsampling matrix, with  $\otimes$  denoting the Kronecker product.

#### 2.4. Hierarchical-based optimization problem

High-dimensional reconstruction algorithms based on hierarchical methodologies aim to decouple a high-complexity optimization problem into a set of low-complexity subproblems. We build a spectral video reconstruction from the proposed temporal windowing sensing methodology by

160 splitting the high-dimensional optimization problem into  $W^2$  low-complexity problems. Based  
 161 on the PnP-ADMM framework [31], the optimization problem can be modeled as

$$\arg \min_{\mathbf{f}_{i_w}, \boldsymbol{\theta}_{i_w}} \|\mathbf{g}_{i_w} - \mathbf{H}_{i_w} \mathbf{f}_{i_w}\|_2^2 + \frac{\mu}{2} \|\boldsymbol{\theta}_{i_w} + \boldsymbol{\eta}_{i_w} - \mathbf{f}_{i_w}\|_2^2 + \phi(\boldsymbol{\theta}_{i_w}), \quad (11)$$

where  $\mu > 0$  is a regularization parameter,  $\boldsymbol{\theta}_{i_w} \in \mathbb{R}^n$  is an auxiliary parameter,  $\phi(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}$  is a prior function that plays the role of an arbitrary regularizer, and  $\boldsymbol{\eta}_{i_w} \in \mathbb{R}^n$  is the scaled dual variable. To improve the reconstruction performance, the optimization problem in Eq. (11) is initialized with a pre-recovered grayscale video version  $\mathbf{z}_{i_w}$  of the scene directly from the measurement  $\mathbf{g}_{i_w}$ . This approach is inspired by image processing methods that have demonstrated that the initial estimations can speed up convergence and avoid local minima [28–30]. Similarly to Eq. (11), the optimization problem to estimate  $\mathbf{z}_{i_w}$  can be written as

$$\arg \min_{\mathbf{z}_{i_w}, \boldsymbol{\omega}_{i_w}} \|\mathbf{g}_{i_w} - \mathbf{H}_{i_w} \mathbf{D}^T \mathbf{z}_{i_w}\|_2^2 + \frac{\mu_z}{2} \|\boldsymbol{\xi}_{i_w} + \mathbf{z}_{i_w} - \boldsymbol{\omega}_{i_w}\|_2^2 + \lambda_z \phi(\boldsymbol{\omega}_{i_w}), \quad (12)$$

162 where  $\mu_z > 0$  is the weighting of the augmented lagrangian term,  $\boldsymbol{\xi}_{i_w} \in \mathbb{R}^{N_x N_y S}$  is the dual  
 163 variable,  $\boldsymbol{\omega}_{i_w} \in \mathbb{R}^{N_x N_y S}$  is an auxiliary variable with  $\boldsymbol{\omega}_{i_w} = \mathbf{z}_{i_w}$ , and  $\phi(\cdot) : \mathbb{R}^{N_x N_y S} \rightarrow \mathbb{R}$  is a  
 164 prior restriction function for which their structure does not need to be defined in PnP-ADMM  
 165 approaches. Since the proposed sensing methodology allows recovering in parallel the SV in  $W^2$   
 166 low-complexity problems, the proposed hierarchical-based approach reduces the computational  
 167 cost from  $O(W^6 n^3)$  to  $O(n^3)$ . The PnP-ADMM procedure is summarized in Algorithm 1. In line  
 168 2, the variables  $\mathbf{z}_{i_w}^{(0)}$ ,  $\boldsymbol{\omega}_{i_w}^{(0)}$ , and  $\boldsymbol{\xi}_{i_w}^{(0)}$  are initialized as all-zeros vectors. The PnP-ADMM iterations  
 169 for the grayscale approximation are computed in lines 4, 5, and 6, where  $\mathcal{L}_{\mu_z}(\cdot)$  corresponds to  
 170 the cost function stated in Eq. (12). In line 8, the variables  $\mathbf{f}_{i_w}^{(0)}$ ,  $\boldsymbol{\theta}_{i_w}^{(0)}$ , and  $\boldsymbol{\eta}_{i_w}^{(0)}$  are initialized by  
 171 using the calculated variables in the first loop. Then, the PnP-ADMM iterations for the spectral  
 172 video are evaluated in lines 10, 11, and 12 here,  $\mathcal{L}_{\mu}$  is the cost function defined in Eq. (11).  
 173 Finally, the recovered video  $\mathbf{f}_{i_w}$  is returned in line 14.

### 174 3. Results

175 The proposed CSV methodology is validated via simulations using a dataset of eleven spectral  
 176 videos. In particular, the first ten videos were created using the RGB-to-spectral mapping  
 177 network [32] over ten selected videos of the *Need for Speed* dataset [33]. Then, after a  
 178 resized/cropping step, the resulting spectral videos have a spatial resolution of  $N_x \times N_y = 720 \times 720$ ,  
 179 a spectral resolution of  $N_\lambda = 8$ , and a temporal resolution of  $N_t = 32$ . The eleventh spectral

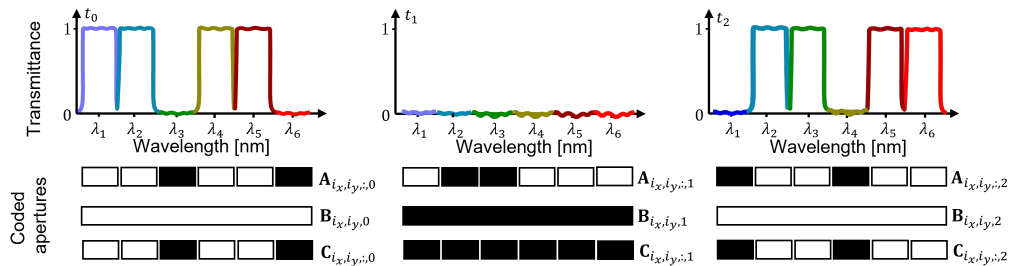


Fig. 3. Illustration of the proposed coding structure for an arbitrary spatial position  $(i_x, i_y)$  at the detector in a spectral video with  $N_\lambda = 6$  spectral bands and  $N_t = 3$  frames.

---

**Algorithm 1** PnP-ADMM spectral video recovery

---

```
1: Input: Acquired data  $\{\mathbf{g}_{i_w}, \mathbf{H}_{i_w}\}$ , and maximum number of iterations  $T_1$  and  $T_2$ 
2: Initialize:  $\mathbf{z}_{i_w}^{(0)} = \mathbf{0}$ ,  $\omega_{i_w}^{(0)} = \mathbf{0}$  and  $\xi_{i_w}^{(0)} = \mathbf{0}$ 
3: for  $t = 1 : T_1 - 1$  do
4:    $\mathbf{z}_{i_w}^{(t+1)} := \arg \min_{\mathbf{z}_{i_w}} \mathcal{L}_{\mu_z}(\mathbf{z}_{i_w}, \omega_{i_w}^{(t)}, \xi_{i_w}^{(t)})$  ▷ Grayscale video
5:    $\omega_{i_w}^{(t+1)} := \arg \min_{\omega_{i_w}} \mathcal{L}_{\mu_z}(\mathbf{z}_{i_w}^{(t+1)}, \omega_{i_w}, \xi_{i_w}^{(t)})$  ▷ Denoising step
6:    $\xi_{i_w}^{(t+1)} := \xi_{i_w}^{(t)} + \mathbf{z}_{i_w}^{(t+1)} - \omega_{i_w}^{(t+1)}$  ▷ Dual update
7: end for
8: Initialize:  $\mathbf{f}_{i_w}^{(0)} = \mathbf{D}^T \mathbf{z}_{i_w}^{(T)}$ ,  $\theta_{i_w}^{(0)} = \mathbf{D}^T \omega_{i_w}^{(T)}$  and  $\eta_{i_w}^{(0)} = \mathbf{0}$ 
9: for  $t = 1 : T_2 - 1$  do
10:   $\mathbf{f}_{i_w}^{(t+1)} := \arg \min_{\mathbf{f}_{i_w}} \mathcal{L}_{\mu}(\mathbf{f}_{i_w}, \theta_{i_w}^{(t)}, \eta_{i_w}^{(t)})$  ▷ Spectral video
11:   $\theta_{i_w}^{(t+1)} := \arg \min_{\theta_{i_w}} \mathcal{L}_{\mu}(\mathbf{f}_{i_w}^{(t+1)}, \theta_{i_w}, \eta_{i_w}^{(t)})$  ▷ Denoising step
12:   $\eta_{i_w}^{(t+1)} := \eta_{i_w}^{(t)} + \theta_{i_w}^{(t)} - \mathbf{f}_{i_w}^{(t+1)}$  ▷ Dual update
13: end for
14: Return: Recovered video  $\mathbf{f}_{i_w}$ 
```

---

180 video was recorded using a monochromator light source, a motorized linear stage, and a camera  
181 sensor [34]. The resulting video was cropped to have the same dimension as the other synthetic  
182 spectral videos. The sensing matrix is constructed following Eq. (8). Note that  $\mathbf{C}$  is generated  
183 from  $\{\mathbf{A}, \mathbf{B}\}$ , and its resulting transmittance  $t_c$  can be calculated as  $t_c = t_a \cdot t_b$ . These two  
184 parameters have important repercussions on the reconstruction performance; thus, these are  
185 studied in detail in the following sections. All simulations were conducted using an Intel Xeon  
186 ES-2697 2.6 GHz processor with 192 GB RAM memory.

### 187 3.1. Reconstruction performance varying $t_a$ , $t_b$ , and $W$

188 We conducted comprehensive studies on how the transmittance of each coded aperture and the  
189 window size would affect the reconstruction performance. For this analysis, the parameters were  
190 set as  $\{t_a, t_b\} \in \{0.25, 0.5, 0.75\}$ , and  $W \in \{1, 2, 3, 4, 5\}$ . Reconstruction results are summarized  
191 in Figure 4 in terms of the peak signal-to-noise ratio (PSNR), structural similarity index metric  
192 (SSIM), spectral angle mapper (SAM), and the customized general perceptual error metric  
193 (GPEM, details are summarized in Appendix 4.1). For each of the used metrics, the overall  
194 best configuration for  $W$ ,  $t_a$ , and  $t_b$  is shown in a continuous green square, and the specific best  
195 pair of  $t_a$ , and  $t_b$  is shown with a dotted orange square. These results show that the optimal  
196 reconstruction performance, according to the GPEM, in the proposed methodology is achieved  
197 when  $t_a = 0.75$ ,  $t_b = 0.25$ , and  $W = 3$ . Conversely and based on GPEM, the worst scenario is  
198 presented when  $t_a = 0.25$ ,  $t_b = 0.25$ , and  $W = 1$ . In summary, the results show that the spatial  
199 reconstruction performance tends to improve for high transmittance values (i.e.,  $t_a \rightarrow 0.75$ ) in  
200 the first and second masks  $\mathbf{A}, \mathbf{B}$ . Conversely, the spectral reconstruction performance tends to  
201 improve for high transmittance values in  $\mathbf{A}$  (i.e.,  $t_a \rightarrow 0.75$ ) and low transmittance values in  $\mathbf{B}$   
202 (i.e.,  $t_b \rightarrow 0.25$ )

203 To further compare the image quality, a representative reconstructed result of the "Kid" and  
204 "Campesina" scenes (with  $t_a = 0.75$  and  $t_b = 0.25$ ) are shown in Fig.5. In particular, for each  
205 scene, a selected frame of the GT is compared with their corresponding frames reconstructed  
206 using five different values of window sizes. The full movie is shown in Video S1. These  
207 reconstructions echo Fig. 4, which supports the positive impact of alleviating the compression

per pixel via the windowing methodology.

### 3.2. Spatio-temporal reconstruction validation

To demonstrate the ability to acquire/estimate the spatial features in the time courses from measurements acquired with the time windowed strategy, see Fig. 6. In particular, we reconstructed a spectral video of a billiard with two moving balls, with their centroids and circumferences being calculated at every frame, as shown in Fig. 6. This analysis is performed for the two windowing scenarios,  $W = 1$  (worse) and  $W = 3$  (optimal), and by setting their optimal pair of  $t_a, t_b$  (green squares in Fig. 4). In Fig. 6, the first row illustrated the circumference prediction for the GT movie. The second and third row shows the reconstruction results, of five equidistant frames, obtained with  $W = 1$  and  $W = 3$ , respectively. Figure 6 shows that without the temporal windowing (i.e.,  $W = 1$ ) methodology, the reconstruction fails to accurately recover the spatial-temporal details, which can be noticed in the detection failure of the billiard balls. In contrast, using the temporal windowing methodology (i.e.,  $W = 3$ ), accurately recovers each billiard ball's size, shape, and position in the entire sequence, which can be noticed in the successful detection. To quantitatively analyze reconstruction, the centroids of all two balls were traced, as seen in Fig. 6 (a). The relative errors of each centroid along the temporal courses in the reconstruction with and without temporal windowing are presented in Fig. 6 (b). The



Fig. 4. Reconstruction performance analysis by varying  $t_a$ ,  $t_b$ , and  $W$  for the spectral videos used in the experiments section 3 (see Visualization 1). Here, the optimal result for each heatmap is highlighted with a dotted orange square. Additionally, the better result per metric is highlighted with a straight green square.


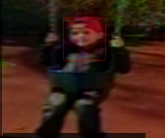



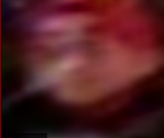





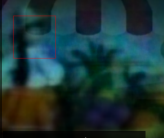
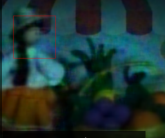


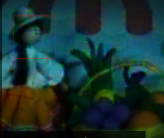

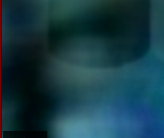




Ground Truth	W = 1	W = 2	W = 3	W = 4	W = 5
					
PSNR / SAM	22.83dB / 16.12°	27.01dB / 14.23°	29.90dB / 12.94°	29.91dB / 12.69°	28.74dB / 13.90°
					
SSIM	0.84	0.89	0.91	0.90	0.89
					
PSNR / SAM	24.25dB / 13.63°	25.65dB / 12.36°	25.77dB / 12.08°	25.14dB / 12.18°	25.06dB / 12.92°
					
SSIM	0.78	0.79	0.79	0.79	0.78

Fig. 5. Visual analysis of the reconstruction of the spectral videos Kid and Campesina using a transmittance of  $\{t_a = 0.75, t_b = 0.25\}$  and varying  $W \in \{1, 2, 3, 4, 5\}$ .

quantitative results show that using temporal windowing produces a higher reconstructed image accuracy than without using the windowing methodology.

### 3.3. Spectral validation

To demonstrate the ability to accurately recover the spectral features, we performed spectral classification of the billard video using a spectral K-means algorithm set up for seven spectral groups [35]. Figure 7 reports the quantitative and qualitative results of this validation, which is performed for the frames  $\{1, 9, 17, 25, 32\}$  of the GT video (first row), and the reconstructed videos using a temporal windowing of  $W = 1$  (second row) and  $W = 3$  (third row). The colors in the segmentation maps show seven representative classes in the scene, comparing with the ground truth SV in Fig. 6 it is possible to relate them to the white and red pool balls, the fabric, and wood of the pool table, and the cue stick. In particular, the classification results attained using  $W = 3$  are 24% and 10% higher in terms of the Accuracy and F1-Score metrics, respectively, than those obtained with  $W = 1$ .

Finally, to further validate the spectral reconstruction accuracy, Fig. 8 illustrates the first four eigenvectors estimated from the spectral covariance matrices for the same spectral frames analyzed in Figs 6-7. In particular, the superiority of the temporal windowing methodology can be appreciated in the estimation of the fourth eigenvector, where the results show an average SAM of  $4.18^\circ$  for  $W = 3$  and  $13.06^\circ$  for  $W = 1$ . These results echo the classification performance presented in Fig. 7. Thus, the temporal windowing methodology leads to a relaxation of the reconstruction problem, which contributes to better rendered image quality.

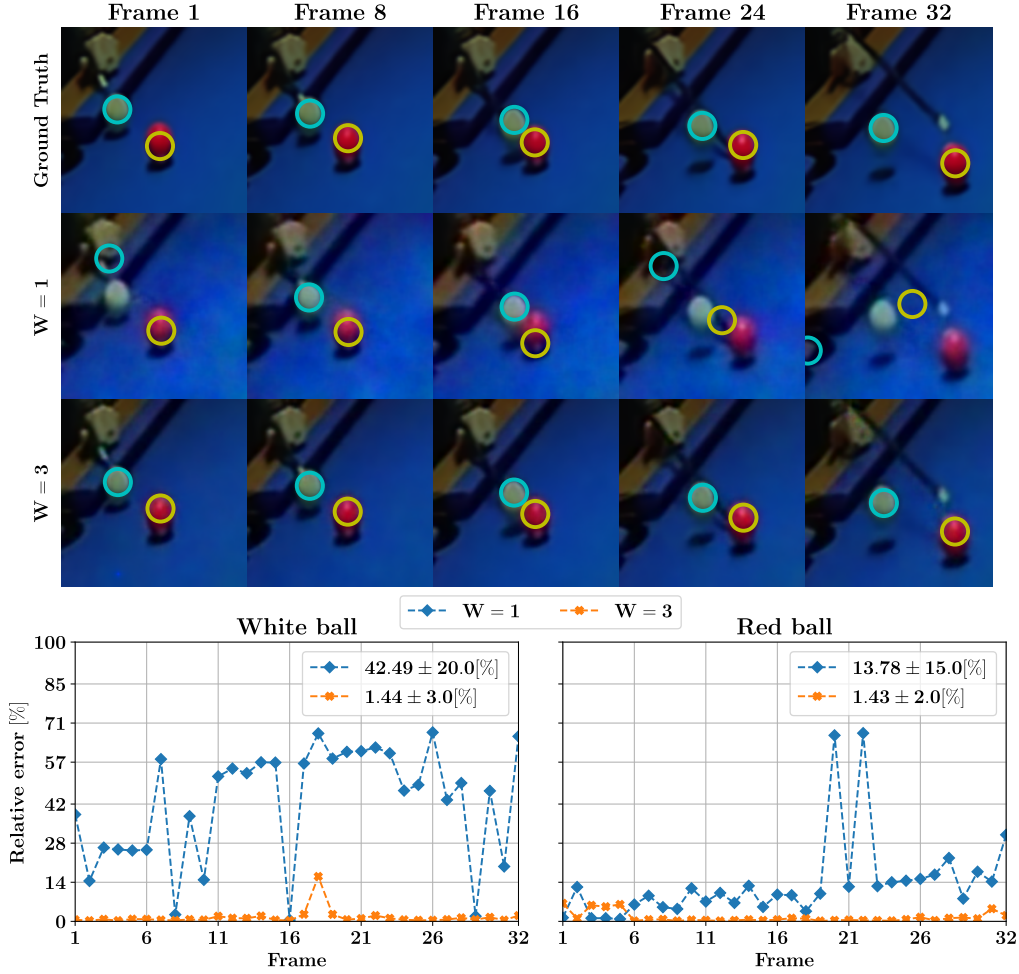


Fig. 6. Results of circle detection experiment over reconstructed billard spectral video using  $W = 1$  and  $W = 3$ , for each pool ball in the scene is calculated the relative error of the predicted centroids.

#### 4. Conclusions

In this article, we propose a CSMI modulation scheme using a dynamic color-coded aperture and windowing approach for wavelength and temporal dimensions, respectively. The encoded spectral and temporal information is acquired within a single image snapshot, where the coded aperture is defined as the multiplication of two independent masks. The proposed coded aperture structure independently modulates the temporal-spectral dimensions, and the proposed windowing approach reduces the compression level, allowing the formulation of a hierarchical reconstruction method, and resulting in the relaxation of the inverse problem. The results suggested that the best spatial-temporal-spectral reconstruction is achieved using a window size  $W = 3$ , and the combination of  $t_a = 0.75$  and  $t_b = 0.25$  transmittance for the masks **A** and **B**. Notice that a high number of window sizes  $W$  increases the temporal resolution while the spatial resolution decreases, then,  $W = 3$  represents a fair trade-off between the spatial and temporal resolution. Furthermore, according to the GPEM, the temporal-spectral recovery performance improves for  $t_a$  and  $t_b$  with high and low values, respectively. The proposed methodology was also compared



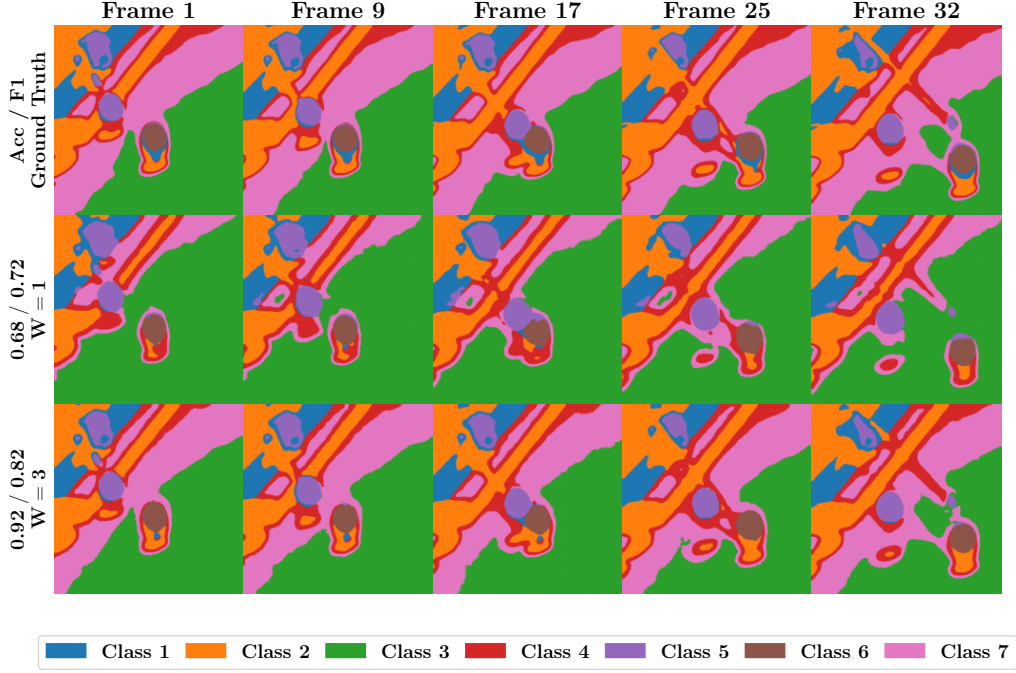


Fig. 7. Results of spectral classification experiment over reconstructed billiard spectral video using  $W = 1$  and  $W = 3$ , for each pixel in the scene, is assigned a class.

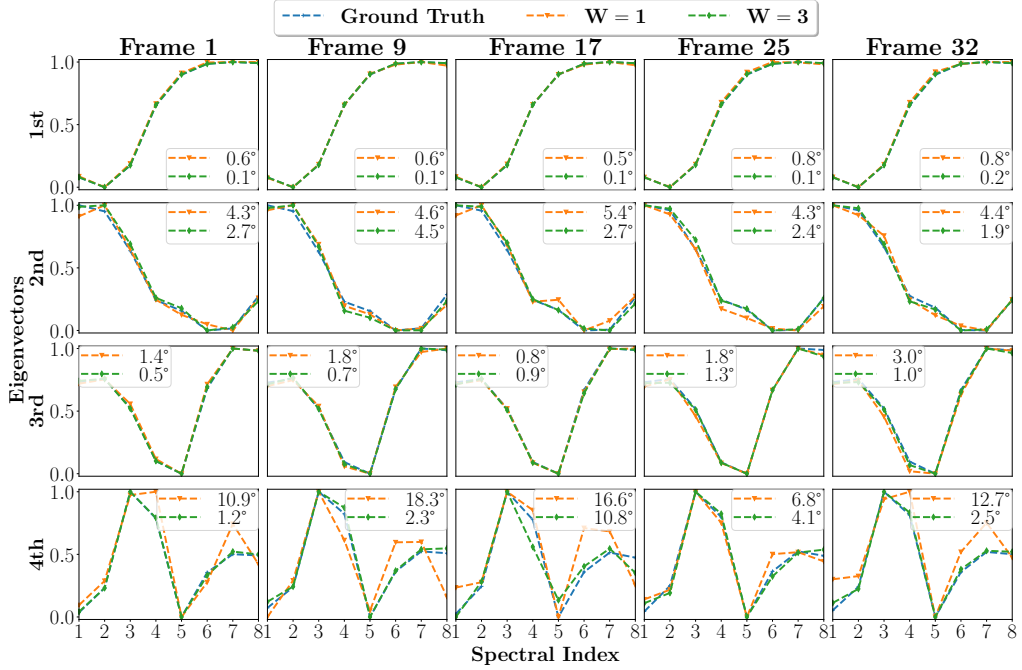


Fig. 8. First four eigenvectors of the covariance matrix for each reconstructed SV using a transmittance of  $\{t_a = 0.75, t_b = 0.25\}$  and  $W \in \{1, 3\}$ . For  $W = 1$  and  $W = 3$  is calculated the SAM of the estimated eigenvectors



in both spectral classification and object detection applications. Results demonstrated superior performance for capturing and recovering key information from compressed spectral video scenes compared to traditional CSV schemes.

## Appendix

### 4.1. General Perceptual Error Metric (GPEM)

Although PSNR, SSIM, and SAM metrics quantify the reconstruction errors, in this work the performance of each metric depends on the scene. Then, a general perceptual error metric (GPEM) is proposed. Similar to [36], the optimal configurations are different for each metric, then, to find the optimal configuration, the GPEM corresponds to a dimensionless normalized linear combination of PSNR, SSIM, and SAM.

$$\tilde{\Phi}_3 = \sum_{k=0}^2 \alpha_k \tilde{\Phi}_k, \quad (13)$$

where,  $\sum_{k=0}^2 \alpha_k = 1$  and  $\hat{\Phi}_k \in \mathbb{R}^{K_1 \times K_2 \times K_3}$  corresponds to the matrix arranging normalized the values of the  $k$ -th metrics (PSNR, SSIM, and SAM) for  $K_1$  videos,  $K_2$  window sizes, and  $K_3$  combinations of transmittances  $t_a$  and  $t_b$ . For the metrics where the greater the better, such as PSNR and SSIM, the normalized metric is defined as  $\hat{\Phi}_k = \frac{\Phi_k - \min(\Phi_k)}{\max(\Phi_k) - \min(\Phi_k)}$ . But for the SAM metric, the lower the better  $\hat{\Phi}_k = 1 - \frac{\Phi_k - \min(\Phi_k)}{\max(\Phi_k) - \min(\Phi_k)}$ . Where  $\max(\cdot)$  and  $\min(\cdot)$  are the maximal and minimal values of each matrix.

## 5. Backmatter

**Funding.** This work was partially funded by Agencia Nacional de Investigacion y Desarrollo (ANID) ANILLOS ATE220022, Fondo Nacional de Desarrollo Científico y Tecnológico FONDECYT 1201081, FONDECYT EXPLORACION 13220234, Project FRO19101 MINEDUC, and the Vicerrectoría de Investigación Extensión at the Universidad Industrial de Santander, Colombia, project VIE-code 3735.

### Disclosures.

The authors declare no conflicts of interest

## References

1. H. V. Nguyen, A. Banerjee, and R. Chellappa, "Tracking via object reflectance using a hyperspectral video camera," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, (2010), pp. 44–51.
2. K. J. Zuzak, S. C. Naik, G. Alexandrakakis, *et al.*, "Intraoperative bile duct visualization using near-infrared hyperspectral video imaging," *The Am. J. Surg.* **195**, 491–497 (2008).
3. R. Leitner, M. D. Biasio, T. Arnold, *et al.*, "Multi-spectral video endoscopy system for the detection of cancerous tissue," *Pattern Recognit. Lett.* **34**, 85–93 (2013).
4. G. Shaw and H. K. Burke, "Spectral imaging for remote sensing," *Linc. laboratory journal* **14**, 3–28 (2003).
5. E. Honkavaara, H. Saari, J. Kaivosoja, *et al.*, "Processing and assessment of spectrometric, stereoscopic imagery collected using a lightweight uav spectral camera for precision agriculture," *Remote. Sens.* **5**, 5006–5039 (2013).
6. P. Kumar, A. Mittal, and P. Kumar, "Fusion of thermal infrared and visible spectrum video for robust surveillance," in *Computer Vision, Graphics and Image Processing*, (Springer, 2006), pp. 528–539.
7. M. Gerken, J. Fritze, M. Münzberg, and M. Weispfenning, "Military reconnaissance platform for the spectral range from the visible to the mwr," in *Infrared Technology and Applications XLIII*, vol. 10177 (International Society for Optics and Photonics, 2017), p. 101770C.
8. G. I. Metternicht and J. Zinck, "Remote sensing of soil salinity: potentials and constraints," *Remote. sensing Environ.* **85**, 1–20 (2003).
9. D. Donoho, "Compressed sensing," *IEEE Trans. on Inf. Theory* **52**, 1289–1306 (2006).
10. A. A. Wagadarikar, N. P. Pitsianis, X. Sun, and D. J. Brady, "Video rate spectral imaging using a coded aperture snapshot spectral imager," *Opt. express* **17**, 6368–6388 (2009).
11. L. Wang, Z. Xiong, H. Huang, *et al.*, "High-speed hyperspectral video acquisition by combining nyquist and compressive sampling," *IEEE transactions on pattern analysis machine intelligence* **41**, 857–870 (2018).

12. X. Ma, X. Yuan, C. Fu, and G. R. Arce, "Led-based compressive spectral-temporal imaging," *Opt. Express* **29**, 10698–10715 (2021).
13. K. M. León-López, L. V. G. Carreno, and H. A. Fuentes, "Temporal colored coded aperture design in compressive spectral video sensing," *IEEE Trans. on Image Process.* **28**, 253–264 (2018).
14. C. A. Barajas-Solano, J.-M. Ramirez, and H. Arguello, "Spectral video compression using convolutional sparse coding," in *2020 Data Compression Conference (DCC)*, (2020), pp. 253–262.
15. C. Barajas-Solano, J.-M. Ramirez, J. I. M. Torre, and H. Arguello, "Compressive spectral video sensing using the convolutional sparse coding framework csc4d," *J. Vis. Commun. Image Represent.* **92**, 103782 (2023).
16. E. J. Candes, "The restricted isometry property and its implications for compressed sensing," *Comptes rendus. Math.* **346**, 589–592 (2008).
17. A. Amini and F. Marvasti, "Deterministic construction of binary, bipolar, and ternary compressed sensing matrices," *IEEE Trans. on Inf. Theory* **57**, 2360–2370 (2011).
18. Y. Mejia and H. Arguello, "Binary codification design for compressive imaging by uniform sensing," *IEEE Trans. on Image Process.* **27**, 5775–5786 (2018).
19. E. Vargas, H. Rueda-Chacón, and H. Arguello, "Learning time-multiplexed phase-coded apertures for snapshot spectral-depth imaging," *Opt. Express* **31**, 39796–39810 (2023).
20. J. Chang and G. Wetzstein, "Deep optics for monocular depth estimation and 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019).
21. H. Arguello, J. Bacca, H. Kariyawasam, *et al.*, "Deep optical coding design in computational imaging: a data-driven framework," *IEEE Signal Process. Mag.* **40**, 75–88 (2023).
22. V. Sitzmann, S. Diamond, Y. Peng, *et al.*, "End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging," *ACM Trans. on Graph. (TOG)* **37**, 1–13 (2018).
23. T.-H. Tsai, P. Llull, X. Yuan, *et al.*, "Spectral-temporal compressive imaging," *Opt. letters* **40**, 4054–4057 (2015).
24. L. Galvis, E. Mojica, H. Arguello, and G. R. Arce, "Shifting colored coded aperture design for spectral imaging," *Appl. optics* **58**, B28–B38 (2019).
25. H. Rueda-Chacon, F. Rojas, and H. Arguello, "Compressive spectral image fusion via a single aperture high throughput imaging system," *Sci. Reports* **11**, 10311 (2021).
26. X. Yuan, J. Yang, P. Llull, *et al.*, "Adaptive temporal compressive sensing for video," in *2013 IEEE International Conference on Image Processing*, (2013), pp. 14–18.
27. Y. Ma, J. Wu, S. Chen, and L. Cao, "Explicit-restriction convolutional framework for lensless imaging," *Opt. Express* **30**, 15266–15278 (2022).
28. S. Marchesini, Y.-C. Tu, and H.-t. Wu, "Alternating projection, ptychographic imaging and phase synchronization," *Appl. Comput. Harmon. Anal.* **41**, 815–851 (2016).
29. L. Valzania, J. Dong, and S. Gigan, "Accelerating ptychographic reconstructions using spectral initializations," *Opt. Lett.* **46**, 1357–1360 (2021).
30. M. Marquez, P. Meza, F. Rojas, *et al.*, "Snapshot compressive spectral depth imaging from coded aberrations," *Opt. Express* **29**, 8142–8159 (2021).
31. X. Yuan, Y. Liu, J. Suo, and Q. Dai, "Plug-and-play algorithms for large-scale snapshot compressive imaging," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), pp. 1447–1457.
32. Y. Zhao, L.-M. Po, Q. Yan, *et al.*, "Hierarchical regression network for spectral reconstruction from rgb images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, (2020), pp. 422–423.
33. H. Kiani Galoogahi, A. Fagg, C. Huang, *et al.*, "Need for speed: A benchmark for higher frame rate object tracking," in *Proceedings of the IEEE International Conference on Computer Vision*, (2017), pp. 1125–1134.
34. K. M. L. López, "Design and optimization of a compressive spectral video sensing system," Ph.D. thesis, Universidad Industrial de Santander (2022).
35. D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," *Tech. rep.*, Stanford (2006).
36. M. Marquez, H. Rueda-Chacon, and H. Arguello, "Compressive spectral imaging via virtual side information," *IEEE Trans. on Comput. Imaging* **7**, 114–123 (2021).