Assembly, fusion and coded aperture design of two compressive spectral imaging sensors via deep learning end-to-end optimization.

Román Alejandro Jácome Carrascal

A Thesis Presented in Fulfillment of the Requirements for the Degree of Electronic Engineer

Advisor:

Ph.D(c) Jorge Luis Bacca Quintero

Co-directors:

Ph.D Henry Arguello Fuentes

Ph.D Julian Rodríguez Ferreira

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones

Bucaramanga

2021

Dedicatoria

A mis padres por todas sus enseñanzas,
amor y apoyo incondicional

A Karen por ser mi compañera de vida
y alegrar mis días con su sonrisa

Agradecimientos

A mis compañeros de carrera que han sido parte de gran importancia en mi vida.

A mi director Jorge Bacca por todas sus enseñanzas, acompañamiento y tiempo dedicado en mi formación.

Al profesor Henry Arguello por brindarme sus consejos y por ofrecerme la oportunidad de ser parte del grupo de investigación HDSP

A los integrantes del grupo HDSP por todo el apoyo brindado en mi formación académica y profesional.

Table of Contents

List of Figures

List of Tables

Resumen

Título: Montaje, fusion y diseño de aperturas codificadas de dos sistemas de adquisición de imágenes espectrales comprimidas por medio de aprendizaje profundo de extremo a extremo [*]

Autor: Román Alejandro Jácome Carrascal [**]

Palabras Clave: Optimización de extremo a extremo, Métodos de fusión, Imágenes Espectrales Comprimidas

Descripción: La adquisición imágenes espectrales compresivas (CSI) reducen la cantidad de datos capturados mediante el uso de proyecciones 2D de la señal 3D original; en consecuencia, es necesario abordar un proceso de recuperación para obtener la señal original. La sistemas CSI sacrifican la resolución espacial para conseguir una alta resolución espectral o viceversa. Por ello, enfoques recientes se basan en la fusión de dos sistemas CSI para obtener una alta resolución espectral-espacial. Los sistemas de detección compresiva suelen tener un conjunto de parámetros físicos, como las aperturas codificadas, que pueden diseñarse para mejorar la calidad de la reconstrucción. El presente proyecto de grado propone un enfoque de aprendizaje profundo de extremo a extremo para diseñar, ensamblar y fusionar dos sistemas CSI, la arquitectura hiperspectral CASSI (coded aperture snapshot spectral imager) con alta resolución espectral y baja resolución espacial y una arquitectura multiespectral de baja resolución espectral y alta resolución espacial. La metodología propuesta consiste en redes neuronales profundas que aprenden una apertura codificada óptima del sistema CASSI y la matriz de filtros de color de la arquitectura multiespectral restringiendo el proceso de aprendizaje a valores implementables, luego una red neuronal profunda desenrollada realiza la fusión de las dos medidas. Para validar los resultados de simulación, estos sistemas se montarán y calibrarán en el laboratorio óptico-electrónico para capturar escenas reales.

---

[*] Trabajo de grado

[**] Facultad de Ingenierías Fisicomecánicas. Escuela de Ingeniería Eléctrica, Electrónica y de Telecomunicaciones. Director: Ph.D(c) Jorge Luis Bacca Quintero.

Abstract

Title: Assembly, fusion and coded aperture design of two compressive spectral imaging sensors via deep learning end-to-end optimization. [*]

Author: Román Alejandro Jácome Carrascal [**]

Keywords: End-to-End Optimization, Fusion Methods, Compressive Spectral Imaging, Deep learning.

Description: Compressive spectral imaging (CSI) reduces the amount of data captured using 2D projections of the original 3D signal; consequently, a recovery process needs to be addressed to obtain the original signal. To date, most CSI sacrifices spatial resolution to achieve high-spectral resolution or vice versa. Therefore, recent approaches are based on the fusion of two CSI systems to obtain a high-spectral-spatial resolution. The compressive sensing systems usually have a set of physical parameters such as coded apertures designed to improve the reconstruction quality. To address this issue, this degree project proposes an end-to-end deep learning approach to design, assemble, and fusion two CSI systems, the hyperspectral CASSI (coded aperture snapshot spectral imager) architecture with high spectral resolution and low spatial resolution and a multispectral patterned architecture with low-spectral resolution and high-spatial-resolution. The proposed methodology consists of deep neural networks which learn an optimal the coded aperture of the CASSI system and the color filter array of the multispectral patterned constraining the learning process to implementable values. An unrolling deep neural network performs the fusion of the two compressive measurements. To validate the simulation results, these systems will be assembled and calibrated in the optical-electronic laboratory to capture real scenes.

---

[*] Bachelor Thesis

[**] Facultad de Ingenierías Fisicomecánicas. Escuela de Ingeniería Eléctrica, Electrónica y de Telecomunicaciones. Director: Ph.D(c) Jorge Luis Bacca Quintero.

Introduction

Spectral imaging (SI) consists of acquiring 2D images or spectral bands across several spectral points of the electromagnetic field, which conforms to a 3D data cube. This information allows estimating unique characteristics and distribution of the different materials in a scene. Hence, spectral image information is valuable in medical applications Li et al. (2016), remote sensing Govender et al. (2007), art conservation Fischer and Kakoulli (2006), among others. Spectral images can be classified into two groups depending on their spatial and spectral resolution; multispectral images, which have high spatial resolution and low spectral resolution, and hyperspectral images, which have a low spatial resolution but high spectral resolution.

Spectral imaging scanning techniques such as whisk-broom Vane et al. (1993), push-broom Gupta and Hartley (1997) or scanning by wavelengthsGat (2000), acquire the 3D data cube by scanning a single point, a line of the scene, or 2D images at specific wavelengths, respectively. These sensing methods are considered low speed at capturing an objective since they only scan a piece of a target, and therefore for capturing the entire scene, it is required to do several shots, which slows down the imaging process Hagen and Kudenov (2013). Moreover, the amount of data acquired with scanning methods is significantly high and increases processing and storage costs.

On the other hand, to lighten the amount of data captured and decrease the scene's imaging process, compressive spectral imaging (CSI) aims to capture simultaneously spatial and spectral information by acquiring a 2D projection of the original 3D objective data cube Arce et al. (2014).

These methods exploit the compressive sensing theory Candes and Wakin (2008) which states that the original signal can be recovered from fewer measurements than the proposed in Nyquist-Shannon theorem Jerri (1977), under the condition that the signal is sparse on some basis Candes and Romberg (2007). The compression ratio in CSI Many CSI architectures have been proposed Hagen and Kudenov (2013) with different spatial and spectral resolution. The CSI process principle is to modulate the spatial and spectral information; a coded aperture usually performs the spatial modulation, and the spectral encoding process is carried out by a dispersive element Wagadarikar et al. (2008). For instance, the Coded Aperture Snapshot Spectral Imager (CASSI) Wagadarikar et al. (2008) exploits a rich spectral resolution with the dispersive element but sacrifice spatial resolution, while another CSI system is the Multispectral Color Filter Array (MCFA) Rueda et al. (2016) which encodes the spatial and spectral information using a color filter array usually obtained high spatial resolution sacrificing spectral bands.

An essential stage in CSI is the recovery algorithm that estimates the original data cube from the compressive measurements, and this is an ill-posed inverse problem because the dimension of the projections is much smaller than the original HSI dimension Foucart and Rauhut (2013). Traditional methods use convex optimization algorithms Bioucas-Dias and Figueiredo (2007); Figueiredo et al. (2007) to minimize the data fidelity error and a regularization term to solve the ill-posed problem. Usually, this term assumes the target's sparsity in a given representation basis Arce et al. (2014). The recent developments of deep learning-based methods to solve inverse problems Ongie et al. (2020) have shown remarkable performance. Consequently, many works have been proposed

into the CS framework Xiong et al. (2017); Li et al. (2017); Qiao et al. (2020). These methods are based on pure deep learning architectures which suffer from a lack of interpretability in the reconstruction process. To address this problem, the unrolling algorithm Monga et al. (2021) has been proposed. This algorithm gives a structured architecture of the deep neural network based on an iterative optimization method. Notably, in CSI, unrolling-based reconstruction networks have been proposed in Wang et al. (2019a, 2020a). Also, it has been proven in Correa et al. (2016); Arguello and Arce (2014) that the design of the coded aperture, instead of being randomly distributed, of the CSI improves the quality of the estimated data cube using criteria such as the restricted isometry property, which entails the optimal conditions on which a correct reconstruction can be achieved and also the incoherence of the sensing matrix. The representation basis is used as coded aperture criteria. In Wang et al. (2019) is proposed a joint coded aperture design and the reconstruction process in a deep learning end-to-end approach.

A limitation of SI is the spatial-spectral resolution which one scarifies one for the other Amro et al. (2011). A highly studied strategy is the fusion of two spectral images with different spatial and spectral resolutions to overcome this problem. For instance HSI and MSI fusion is one of the most common example of fusion Wei et al. (2015); Yokoya et al. (2012); Gomez et al. (2001). Compressive Spectral Image Fusion (CSIF) has also been studied in Vargas et al. (2018); Gelvez and Arguello (2020); Vargas et al. (2017) to reconstruct a high spatial-spectral resolution from two CSI systems with different encoding strategies. Moreover, deep learning approaches have proposed to solve the fusion issue in Xie et al. (2019); Yang et al. (2018a).

Consequently, in this proposal, we considered a deep learning end-to-end optimization approach for the design and fusion of two CSI systems, one with high spectral resolution and low spatial resolution and the other high spatial resolution and low spectral resolution, to obtain a high spectral and spatial resolution. To design the CSI system, those have to be modeled as a layer of a neural network, leading to a differentiable parameterization of the sensing matrices constraining the learning process to implantable values and an unrolling-based network that performs the fusion of the two compressive measurements.

1. Objectives

General objective

- To assemble, fuse and design the coded apertures of two opto-electronics systems, a hyperspectral CASSI (coded aperture snapshot spectral imager), and a multispectral patterned architecture using a deep learning end-to-end optimization.

Specific objectives

- To model mathematically two spectral imaging systems using a derivable parameterization of the coded aperture and the color filter array to be incorporated as a layer in a neural network under implementable constraints;
- To design and implement a deep neural network architecture to the fusion of the CASSI and patterned measurements. Compare the obtained results with state of the art methods.
- To assemble and calibrate a prototype in the optical laboratory for the fusion of the two optoelectronic systems employing the trained designs.
- To evaluate the performance of the fusion system with real data

2. Spectral Imaging

The spatial and spectral information provided by the spectral images allows the characterization of the different materials in a scene, which have been used to perform important computer vision tasks such as object classification Jácome et al. (2021), image segmentation Camps-Valls et al. (2014), salient object detection Zhang et al. (2018) and object tracking Hien Van Nguyen et al. (2010). These tasks have been widely applied in several fields like medical applications Li et al. (2016), earth observation Datcu and Seidel (2005), food quality Qin et al. (2013) and surveillance Denman et al. (2010) among others. Traditional spectral imaging sensors scan portions of the spatial-spectral scene in each detector integration period and then collect them into a single data cube. For instance, line scanning systems such as the whisk-broom system Vane et al. (1993) which captures only a point or pixel of the scene with its respective spectral signature and then scan all the desired pixels, push-broom sensor Gupta and Hartley (1997) acquires a line of the scene with its respective spectral signature which results in 2D images which then are stacked to conform to the 3D data cube. Another traditional sensors are the tunable filter imagers Gat (2000). These sensors capture the entire spatial dimension at a specific wavelength in each detector integration period, the filters vary to obtain the number of spectral bands desired, and then all the 2D images are stacked in the spectral dimension to create the 3D data cube. The main drawbacks with these methods are the large amount of data acquired and the imaging time, which increases proportionally to the desired spatial and spectral resolution. On the other hand, snapshot imaging systems capture the full spatial and spectral information in a single detector integration period. Figure 1

shows the comparison of the measurements acquired with the scanning and snapshot methods.

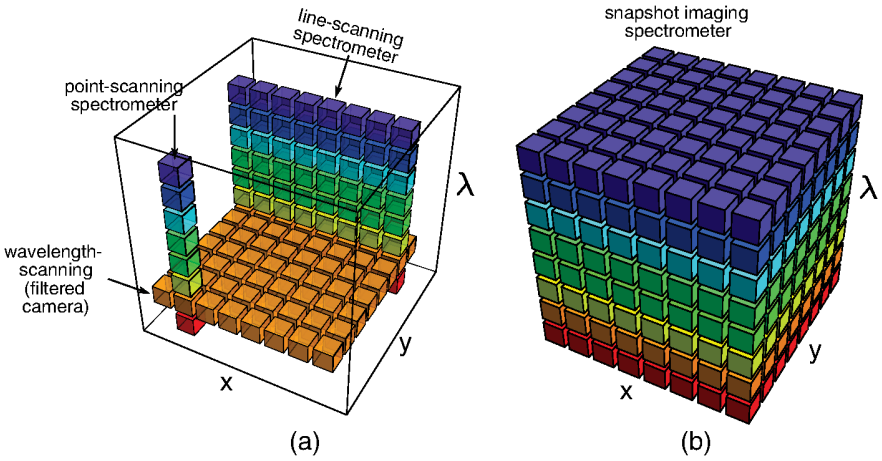

Figure 1. Measurement acquired in single detector integration period for scanning methods (left) and snapshot system (right). Taken from Hagen and Kudenov (2013)

2.1. Compressive Spectral Imaging

To address the large amount of data acquired by traditional SI sensors, CSI theory aims to compress the spatial and spatial data directly from the acquisition process. This approach is inspired by the developments of the compressive sensing (CS) theory Candes and Wakin (2008). CS dictates that it is possible to recover a signal from fewer samples than those required in traditional sampling theories Jerri (1977) if the signal is sparse and the sensing meets with incoherence. The first one stands that the signal is sparse on some proper basis, and the second one refers that the sensing matrix and the representation basis of the signal have low correlation. Candes and Romberg (2007). In CSI, the sparsity in the spectral images has been proved using the Kronecker product of wavelet and cosine transform Arguello and Arce (2014), and the incoherence related to the sensing process has been demonstrated in Arguello and Arce (2012) for CASSI. The sensing process in CSI can be modeled as a linear matrix multiplication: $\mathbf{y} = \mathbf{H}\mathbf{f} + \omega$, where $\mathbf{f} \in \mathbb{R}^{MNL}$ is

the vectorization of the original data cube, M and N are spatial dimensions and L is the number of spectral bands, **H** is the sensing matrix, **y** is the compressive measurements, ω is the noise of the sensor. The matrix **H** and its dimension changes depending on the CSI architecture, likewise the dimension of the compressive measurements **y**. The next subsections will describe two CSI architectures employed for the fusion proposal; CASSI, which allows a high spectral resolution, and MCFA, which allows high spatial resolution.

2.1.1. CASSI. The CASSI architecture is composed of three main elements: The first one is coded aperture such as spatial light modulator (SLM) such as a digital micromirror device (DMD) Galvis et al. (2015) or a liquid crystal SLM Zhu et al. (2013), a dipersive element, a prism or a diffractive grating, and a focal plane array (FPA). For an input light field $f(x,y,\lambda)$, where (x,y) are the spatial coordinates and λ stands for the spectral coordinate, a coded aperture $T_c(x,y)$ encodes the spatial information by blocking or unblocking some pixels of the scene, then, the coded field is sheared into a spatial axis by the dispersive element, which then impinges onto a FPA detector. A continuous CASSI model can be expressed as,

$$y_c(x,y,\lambda) = \int_{\Lambda}\iint T_c(x',y')f(x',y',\lambda)h(x'-x-S(\lambda),y'-y)dx'dy'd\lambda, \tag{1}$$

Λ is the spectral range sensitivity of the detector, $h(\cdot)$ is the optical impulse response, $S(\lambda)$ represents the dispersion function of the dispersive element. Traditional discretization modeling of the CASSI system assumes that one pixel on the FPA detector is just impinged by one sheared voxel; however, authors in Arguello et al. (2013) propose a high-order CASSI model which dictates

that one sheared voxel affects up to three neighboring pixels on the FPA, the modeling of this effect is addressed in the calibration process. Following the modeling in Arguello et al. (2013), a voxel of the input light source is given by

$$F_{i,j,k} = \iiint f(x,y,\lambda) \times \text{rect}\left(\frac{x}{\Delta_{hs}} - i, \frac{y}{\Delta_{hs}} - j, \frac{\lambda}{\Delta_{h\lambda}} - k\right) dxdyd\lambda, \tag{2}$$

where $i, j = 1, ..., M, , j = 1, ..., N$ are spatial indexes and $k = 1, ..., L$ is the spectral index, Δ_{hs} is the pitch size of the high spatial resolution detector and $\Delta_{h\lambda}$ is the spectral high spectral resolution. The coded aperture can be expressed as

$$T_c(x,y) = \sum_{i,j} T_{c_{i,j}} \times \text{rect}\left(\frac{x}{\Delta_{ls}} - t, \frac{y}{\Delta_{ls}} - q\right), \tag{3}$$

where Δ_{ls} is the low spatial resolution sensor pitch. We assume that $\Delta_{ls} > \Delta_{hs}$. Therefore, the discrete version of the measurements are given by

$$\begin{aligned} y_{c_{i,j}} = \iiint f(x + S(\lambda), y, \lambda) \times \text{rect}\left(\frac{x}{\Delta_{hs}} - i, \frac{y}{\Delta_{hs}} - j, \frac{\lambda}{\Delta_{h\lambda}} - k\right) \\ \times \sum_{i,j} T_{c_{i,j}} \times \text{rect}\left(\frac{x}{\Delta_{ls}} - i, \frac{y}{\Delta_{ls}} - j\right) dxdyd\lambda. \end{aligned} \tag{4}$$

Notice that Δ_{ls} and Δ_{hs} can be related as $\Delta_l = r_s \Delta_h$ where $r_s \geq 1$ is a integer which stands for up-sampling factor. The resolution difference between the input source and the coded aperture,

equation (13) is affected as

$$y_{c_{i,j}} = \sum_{k} \iiint \sum_{t,q} f(x+S(\lambda),y,\lambda) \times \text{rect}\left(\frac{x}{\Delta_{hs}} - i - t, \frac{y}{\Delta_{hs}} - j - q, \frac{\lambda}{\Delta_{h\lambda}} - k\right) \times \sum_{i,j} T_{c_{i,j}} \times \text{rect}\left(\frac{x}{\Delta_{ls}} - i, \frac{y}{\Delta_{ls}} - j\right) dxdyd\lambda, \tag{5}$$

where the variables t, q are bounded as

$$\begin{aligned} 1 + (i-1)r_s \leq t \leq ir_s \\ 1 + (j-1)r_s \leq q \leq jr_s, \end{aligned} \tag{6}$$

therefore, the discrete model of (1.1) is depicted as

$$y_{c_{i,j}} = \sum_{k=0}^{L-1} T_{c_{i,j}} \sum_{t=1+(i-1)r_s}^{ir_s} \sum_{t=1+(i-1)r_s}^{ir_s} F_{i-t,j-q-k,k}, \tag{7}$$

where, the shifting in the voxel indexes is due to the dispersion function $S(\lambda)$. The high-order model define three regions $R_0.R_1, R_2$ on the sheared voxel which will impinges one FPA pixel. The energy distribution due to this effect is modeled the weights $w_{i,j,k,u}$ where $u = 0, 1, 2$ stands for the region of the model, see Figure 2. Then, discrete high-order CASSI model is given by

$$y_{c_{i,j}} = \sum_{k=0}^{L-1} \sum_{u=0}^{2} T_{c_{i,j}} \sum_{t=1+(i-1)r_s}^{ir_s} \sum_{2=1+(j-1)r_s}^{jr_s} w_{i,j,k,u} F_{i-t,j-q-k-u,k}. \tag{8}$$

The previous discrete model can be expressed in a matrix-vector product yielding in the following

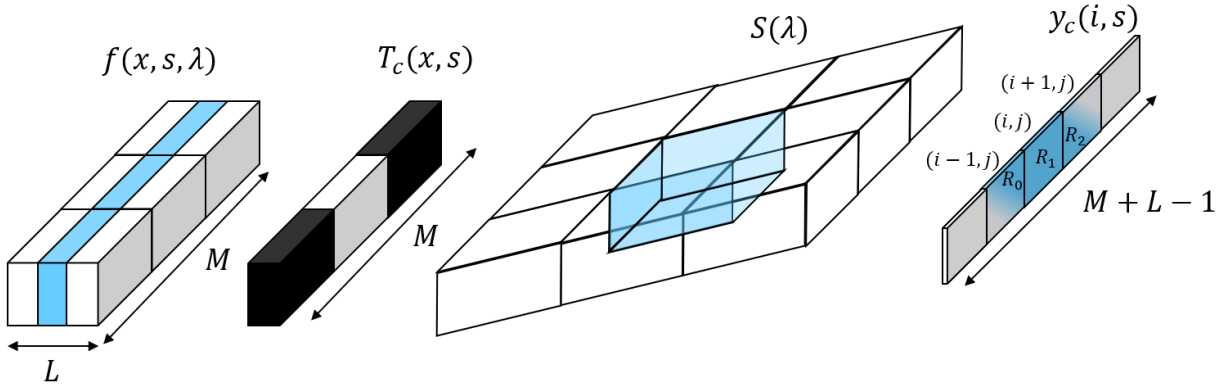

Figure 2. Illustration of the high order modeling of CASSI. For a slice s of the input source, a single voxel impinges up to 3 pixel in the detector

expression

$$\mathbf{y_c} = \mathbf{H}_c\mathbf{f} + \omega_c, \tag{9}$$

where $\mathbf{f} \in \mathbb{R}^{MNL}$ is the vecotrization of the high spatial-spectral resolution image, denote $M_d = \frac{M}{r_s}$ and $N_d = \frac{N}{r_s}$ the down-sampled version of the spatial dimensions, $\mathbf{y}_c \in \mathbb{R}^{M_d(N_d+L-1)}$ is the compressed measurements, the sensing matrix $\mathbf{H_c} \in \mathbb{R}^{M_d(N_d+L-1)K\times MNL}$ which is composed as:

$$\mathbf{H}_c = \mathbf{P}\mathbf{T}_c\mathbf{D}_c, \tag{10}$$

where $\mathbf{D}_c \in \mathbb{R}^{M_dN_dL\times MNL}$ is the spatial decimation operator which down-sample the dimension of the measurements by the factor r_s, $\mathbf{P} \in \mathbb{R}^{M_d(N_d+L-1)\times M_dN_dL}$ models the dispersion effect considering the high-order CASSI modeling and, $\mathbf{T}_c \in \{0,1\}^{M_d(N_d+L-1)\times M_dN_dL}$ diagonalized coded aperture entries $\mathbf{t}_c \in \{0,1\}^{M_dN_d}$. The structure of the sensing matrix $\mathbf{H}_c$ for $M = N = 6$ and, $L = 3$. Notice that the three diagonal structures of $\mathbf{H}_c$ is due to the three regions of the high-order model,

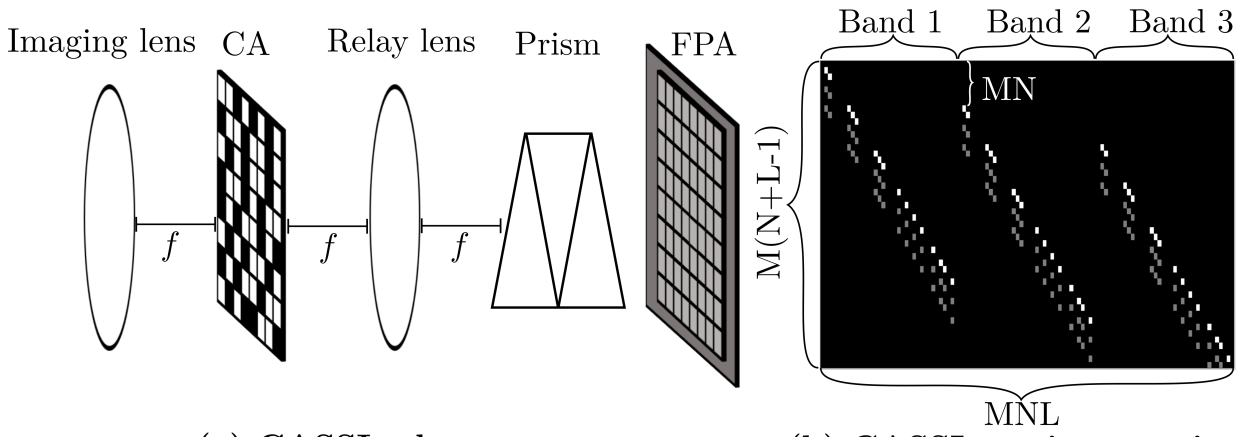

Figure 3. Scheme of CASSI optical architecture (a). Structure of the high-order sensing matrix for $M = N = 6$ and $L = 3$

and their intensities depend on the weights $w_{i,j,k,u}$.

2.1.2. MCFA. The idea behind this CSI architecture is an extension of the commercial color imaging sensors which use a color filter pattern, such as the Bayer pattern Bayer (1976), particularly, this pattern uses three different filters which are related to the red, green, and blue spectral responses and then performs a demosaicing process to reconstruct the three channels RGB image Kimmel (1999). Predictably, to acquire more than three channels, more different filters have to be used in the array. The important aspect of this sensing method is that the spatial and spectral information is encoded simultaneously, the first one by the spatial distribution of the color filters, the second through the spectral responses of the filters themselves. For this purpose, many color filter arrays have been proposed Lapray et al. (2014). Then the coded field is integrated into an FPA detector (see Figure. 4(a)). A continuous model of the sensing process is given by

$$y_m(x,y) = \int_{\Lambda} \iint T_m(x',y',\lambda) f(x',y',\lambda) h(x'-x,y'-y) dx' dy' d\lambda, \tag{11}$$

where $T_m(x,y,\lambda)$ is the color filter array, $f(x,y,\lambda)$ is the input light source and $h(\cdot)$ is the corresponding optical impulse response. The color filter array can be expressed in terms of its discretized version as

$$T_m(x,y,\lambda) = \sum_{i,j,k} T_{m_{i,j,k}} \times \text{rect}\left(\frac{x}{\Delta_{hs}} - j, \frac{y}{\Delta_{hs}} - i, \frac{\lambda}{\Delta_{l\lambda}} - k\right), \tag{12}$$

where $\Delta_{l\lambda}$ is the low spectral resolution factor, and using the voxel description of the input light source in (2), the discrete expression of the measurements is given by

$$\begin{aligned} y_{m_{i,j}} = & \sum_k \iiint f(x,y,\lambda) \times \text{rect}\left(\frac{x}{\Delta_{hs}} - i, \frac{y}{\Delta_{hs}} - j, \frac{\lambda}{\Delta_{h\lambda}} - k\right) \\ & \times \sum_{i,j} T_{m_{i,j}} \times \text{rect}\left(\frac{x}{\Delta_{hs}} - i, \frac{y}{\Delta_{hs}} - j, \frac{\lambda}{\Delta_{l\lambda}} - k\right) dxdyd\lambda, \end{aligned} \tag{13}$$

considering the effect of the different spectral resolution of the color coded aperture and the input source given by $\Delta_{l\lambda} = r_\lambda \Delta_{h\lambda}$ where the integer $r_\lambda \geq 1$ is the spectral Up-sampling factor. Then, the discrete measurements on the FPA detector is given by

$$y_{m_{i,j}} = \sum_{k=0}^{\frac{L}{r_\lambda}} T_{m_{i,j,k}} \sum_{p=1+(k-1)r_\lambda}^{kr_\lambda} F_{i,j,k-p}, \tag{14}$$

which can be expressed in a matrix-vector product such as

$$\mathbf{y}_m = \mathbf{H}_m \mathbf{f} + \omega_m, \tag{15}$$

where $\mathbf{y}_m \in \mathbb{R}^{MN}$, $\omega_m \in \mathbb{R}^{MNL}$ is the noise in the sensing process and $\mathbf{H}_m \in \mathbb{R}^{MN \times MNL}$ is the sensing matrix which is composed as

$$\mathbf{H}_m = \mathbf{T}_m \mathbf{D}_m, \tag{16}$$

denote $L_d = \frac{L}{r_\lambda}$ is the scaled version of the spectral dimension, such as $\mathbf{D}_m \in \{0,1\}^{MNL_d \times MNL}$ is the spectral decimation operator which induces a the factor r_λ in the spectral dimension, $\mathbf{T}_m \in \{0,1\}^{MN \times MNL_d}$ is the diagonalization of the color filter array entries $\mathbf{t}_m \in \{0,1\}^{MNL_d}$. The structure of $\mathbf{H}_m$ is shown in Figure. 4(b) for $M = N = 6$ and $L = 3$

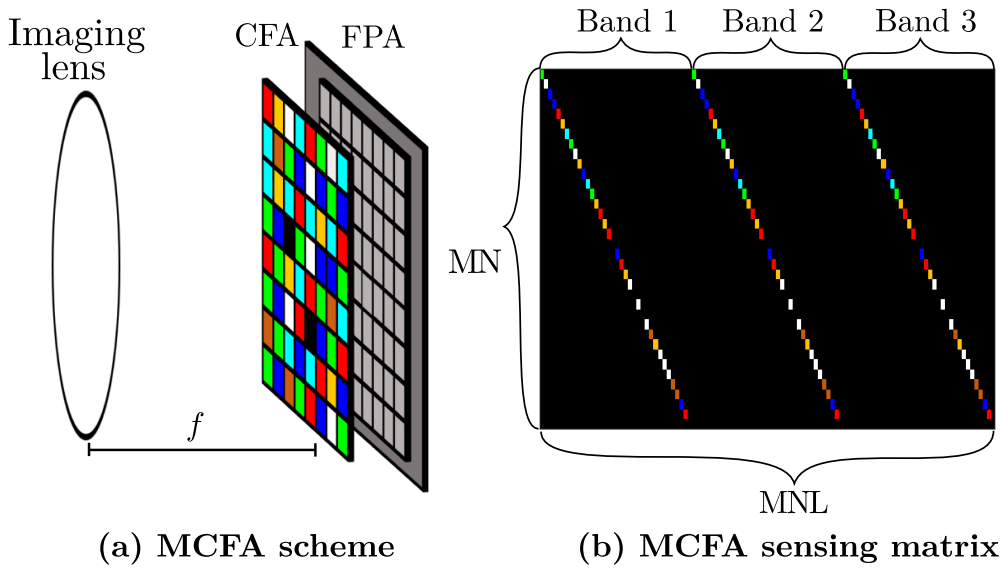

Figure 4. a) Scheme of the MCFA acquisition system. (b) Sensing matrix of a MCFA with $M = N = 6$ and $L = 3$

3. CSI Fusion Reconstruction Process

The reconstruction process after obtaining the compressed measurements is a crucial step in the CS framework. The high-dimensional signal can be recovered by solving an optimization problem. Several recovery methods have been proposed, which can be divided into two main groups, traditional reconstruction methods, and deep learning-based approaches.

3.1. Traditional reconstruction algorithms

These approaches aim to solve the recovery problem by solving an optimization optimization. The reconstruction process is an ill-posed inverse problem since $\mathbf{f}$ is greater than the dimension of $\mathbf{y}$, which entails an undetermined linear system. Therefore, a regularization function is used to constraints the solution using prior information of the signal. For instance, spectral images are sparse in Kronecker basis with a 2D wavelet Symmlet-8 basis and the 1D discrete cosine transform (DCT) Arce et al. (2014). Consequently, the optimization for a compressive measurement is given by

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}} ||\mathbf{H}\mathbf{f} - \mathbf{y}||_2^2 + \tau R(\mathbf{f}), \tag{17}$$

where $R(\cdot)$ is the regularization function, for instance, the function $R(\cdot)$ is the ℓ_1 norm of the representation coefficients given a sparse transformation. Algorithms based on this approach with sparsity prior are Figueiredo et al. (2007); Bioucas-Dias and Figueiredo (2007); Beck and Teboulle (2009); Afonso et al. (2010). Other prior information for spectral imaging are low-rank Gelvez

et al. (2017) and total variation Yuan (2016) among others. For the CSI fusion reconstruction problem, the optimization problem can be formulated as

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}} ||\mathbf{H}_m\mathbf{f} - \mathbf{y}_m||_2^2 + ||\mathbf{H}_c\mathbf{f} - \mathbf{y}_c||_2^2 + \tau R(\mathbf{f}), \tag{18}$$

For instance, in Vargas et al. (2018), equation (18) is solved using an Alternating Direction Method of Multiplier(ADMM) algorithm Boyd et al. (2011) to fuse the compressive measurements using the CASSI and a Colored-CASSI Arguello and Arce (2014), Gelvez and Arguello (2020) proposes an ADMM algorithm based on non-local low-rank abundance prior using multispectral and hyperspectral projections in the CASSI system.

3.2. Deep Learning-Based Algorithms

With the recent developments in deep learning for computer vision tasks Guo et al. (2016), and specifically in solving inverse problems in imaging Ongie et al. (2020), several works have been proposed to solve (17) using deep convolutional neural networks (CNN). These methods leverage many public datasets to increase their performance and generalization, and the inference of the trained model has lower computational complexity than the traditional recovery algorithms. Here the optimization formulation is given by

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{K} \sum_{k=0}^{K-1} \mathscr{L}(\mathscr{M}_\theta(\mathbf{H}\mathbf{f}_k), \mathbf{f}_k), \tag{19}$$

where $k = 1, \ldots, K$ are indexes of the training dataset, $\mathscr{L}(\cdot)$ is the loss function, $\mathscr{M}_\theta$, represents the neural network architecture with trainable parameters θ, which are updated in the

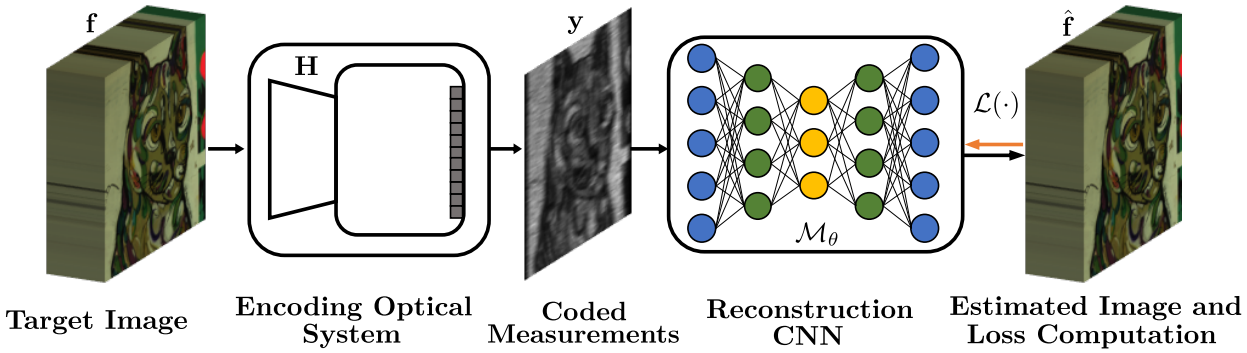

Figure 5. Deep Learning approach for inverse problems, where the input of a CNN is the coded measurements of a target scene, and the output is an estimation of the target image. The network's parameters θ are updated by computing the loss of the estimated image and the ground truth target image. Black arrows represent the forward pass, and orange arrows represent the backpropagation process of the training

back-propagation process, see Figure 5 for a schematic representation. Some approaches have been proposed in this framework. For instance, authors in Mousavi and Baraniuk (2017) proposed a CNN to learn the inversion from the compressive measurements into the target signal through a CNN. In Miao et al. (2019) is proposed a two-stage CNN, a refinement network composed of a 3D U-Net Ronneberger et al. (2015) and a reconstruction stage which uses self-attention modules. Despite the remarkable results obtained by these and other works using CNN, there is a lack of interpretability and flexibility since CNN works as a black box. Then, a combination of these methods and the traditional recovery algorithms have been proposed. For example, the Plug and Play algorithm proposed in Yuan et al. (2020) uses denoising priors to solve the optimization problem. The prior is a pre-trained CNN for the denoising task Tian et al. (2018).

Under these ideas of bringing more flexibility and interpretability to the deep neural network, the unrolling algorithms have attracted enormous attention in several fields to solve ill-posed

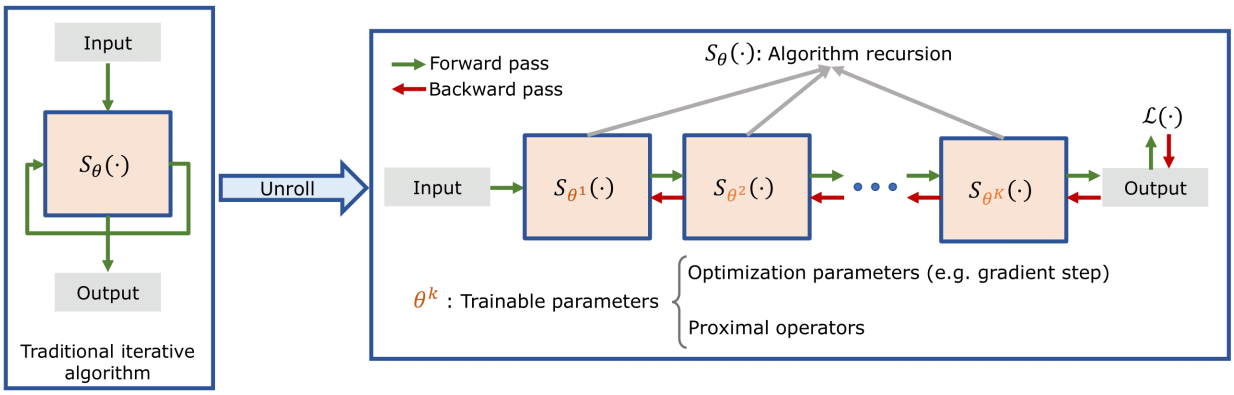

Figure 6. Overview the unrolling algorithm, in which an iterative (left) algorithm is 'unrolled' into a structured neural network where each stage of the network is iteration and in every stage there are trainable parameters that are learned in the backpropagation process from the training dataset.

inverse problems. This framework's main idea is to give structure to a deep neural network based on the iterations of an optimization algorithm. This methodology was first proposed in Gregor and LeCun (2010) where the unrolling algorithm solves a sparse coding problem, but recently it has been proposed to several areas such as speech recognition, medical imaging, and remote sensing Monga et al. (2021). Figure. 6 shows a visual overview of the unrolling algorithm. Specifically, trying to solve (17) in traditional algorithms, requires to performs proximal operators Parikh and Boyd (2014) in each iteration. These operators are hand-crafted selected depending on the prior chosen in the algorithm, but in the unrolling algorithm, this operator can be learned using CNN. Also, optimization parameters chosen through cross-validation or deduced analytically, such as regularization parameters or gradient steps, can be learned in the training process, yielding in an end-to-end training of the recovery algorithm. For CSI, several works have adopted this framework. For instance, authors in Wang et al. (2019b) proposed an unrolled network based on a Half Quadratic Splitting (HQS) formulation of (17) and with a deep spatial and spectral sub-network

prior, in Wang et al. (2020b) the sub-network prior exploits non-local structures and in Sogabe et al. (2020) the unrolling network is based in an Alternating Direction Method of Multipliers (ADMM) algorithm Boyd et al. (2011) and using the same deep spatial and spectral prior of Wang et al. (2019b).

A deep learning formulation for the CSI fusion problem can be expressed as

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{K} \sum_{k=0}^{K-1} \mathscr{L}(\mathscr{M}_{\theta}(\mathbf{H}_m \mathbf{f}_k, \mathbf{H}_c \mathbf{f}_k), \mathbf{f}_k), \tag{20}$$

For, CSI fusion it has not been proposed works in the deep learning approach, but for multispectral and hyperspectral image fusion, there have been several proposals, for instance, in Palsson et al. (2017) the fusion is based on a 3D CNN architecture, and Principal Component Analysis (PCA) dimensionality reduction, Zhou et al. (2019) uses an encoder network for the HS image and concatenate the MS image in the latent space of the network, Wang et al. (2020c) proposed a joint probabilistic generative network composed in a spectral generative network, a spatial-dependent prior network, and a spatial-spectral variational inference network and Yang et al. (2018b) uses two deep branches, one extract spatial features from the MS image and the other extracts features from the spectral data of the HS image, both features are concatenated and fused using fully connected layers. Finally, unrolling methods have also been used in this field. For example, in Xie et al. (2020) is proposed an unrolled which takes into account the observation model and encourage low-rankness of the HS image and they also consider a blind methodology and in Wang et al. (2019) the network earns the observation model, and an iterative refinement unit

is used in each optimization stage where features extracted in each step are concatenated at the end of the network.

Due to the remarkable results in the works aforementioned, in this work, it was decided to adopt the unrolling framework to reconstruct the data cube from the two compressive measurements.

3.3. Coded Aperture Design

Along with the enormous effort in developing efficient algorithms to solve (17) in CSI problems, several works have shown the distribution of the coded aperture affects the quality of the estimated 3D signal. Several design criteria have been used to optimize the coded aperture, the work in Correa et al. (2016) uses the restricted isometry property (RIP) as design criteria, this is a crucial property in CS since it determines the minimum number of measurements needed for correct recovery of the signal, in Arguello and Arce (2014) uses the RIP criteria to optimize the cut-off frequencies of colored coded apertures, Mejia and Arguello (2018) optimizes the coded aperture by minimizing its zero singular values and keeping uniform the number the non-zero entries per column and row. The coded aperture design boosts the recovery process and improves the performance in other computer vision tasks such as classification Hinojosa et al. (2018); Ramirez et al. (2014).

3.4. End-to-End Optimization

A recent trend in computational imaging is the sensing process's joint learning and a decoder operator to recover the original signal denote as an End-to-End (E2E). The E2E approach can be seen as a DNN where the first layer simulates the forward model of the optical sensing

process where some parameters are set as trainable parameters (*e.g.* coded aperture), constraining the training of these parameters into an implementable feasible set such as binary values for coded aperture and, the rest of the network is the decoder operator (*e.g.* CNN, unrolling network, or others) for a specific task. In each training pass, the decoder weights are updated along with the sensing trainable parameters, therefore obtaining an optimal sensing parameter design by minimizing the network's cost function. See Figure 7. Several works have proposed an E2E approach where optical parameters to be optimized are the height maps of a diffractive lens to 3D object detection Chang and Wetzstein (2019), monocular depth estimation and super-resolution Sitzmann et al. (2018), and snapshot spectral imaging and depth estimation Baek et al. (2021). Moreover, these learning capabilities have also been used for optimal coded aperture design Bacca et al. (2020) for classification task and CSI in Wang et al. (2019) is proposed a joint coded aperture optimization and reconstruction, where the weights of the sensing layer are the value of the coded aperture and a binarization function is employed to obtain values of either 0 or 1 for implementation purpose. The formulation of an E2E CSI can be expressed as

$$\{\hat{\mathbf{H}}, \hat{\theta}\} = \underset{\{\mathbf{H}\in\mathscr{R},\theta\}}{\arg\min} \frac{1}{K} \sum_{k=0}^{K-1} \mathscr{L}(\mathscr{M}_{\theta}(\mathbf{H}\mathbf{f}_k), \mathbf{f}_k), \tag{21}$$

where $\mathscr{R}$ is the feasible set for the sensing matrices due to implementation purpose.

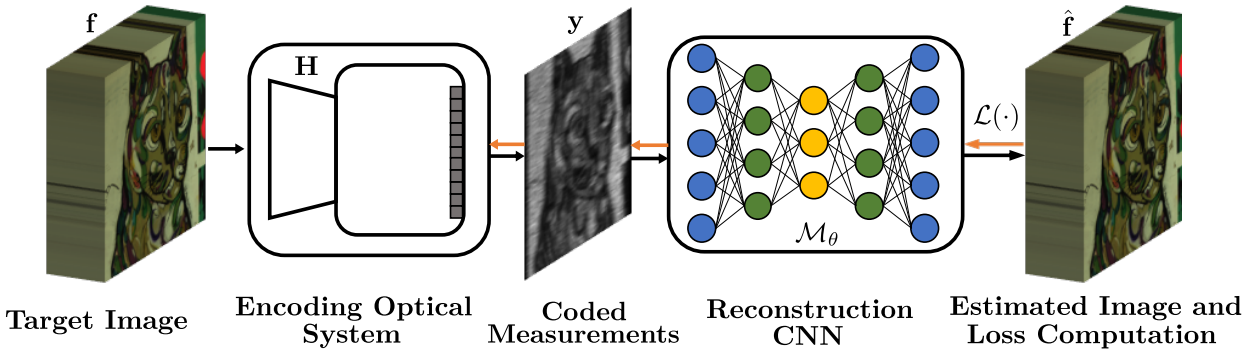

Figure 7. E2E approach reconstruction task. The optical system is jointly updated with the CNN parameters in each training step for E2E learning of the computational imaging system. Black arrows represent the forward pass and orange arrows represent the backpropagation process of the training

4. Unrolling E2E Optimization for CSI Fusion

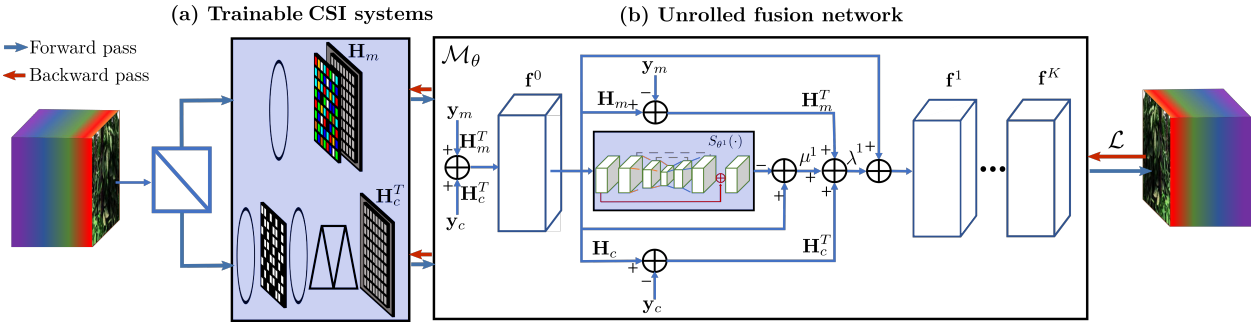

Figure 8. Overview of the proposed E2E unrolled network for CSI fusion. (a) The CSI systems are modeled as NN layer where the customizable (CFA in MCFA and CA in CASSI) are updated in each backward pass according to the loss function values. (b) the proposed optimization inspired fusion network for K iterations.

The E2E formulation for CSI fusion with CASSI and MCFA can be expressed as

$$\{\hat{\mathbf{H}}_m, \hat{\mathbf{H}}_c, \hat{\theta}\} = \underset{\{\mathbf{H}_m \in \mathscr{R}_m, \mathbf{H}_c \in \mathscr{R}_c, \theta\}}{\arg\min} \frac{1}{K} \sum_{k=0}^{K-1} \mathscr{L}(\mathscr{M}_\theta(\mathbf{H}_m \mathbf{f}_k, \mathbf{H}_c \mathbf{f}_k), \mathbf{f}_k), \tag{22}$$

where $\mathscr{R}_m$ and $\mathscr{R}_c$ are the feasible sets for the MCFA and CASSI systems. This constraints will be detailed in the following section

4.1. Layer Modeling of the Optical Systems

For an E2E approach, the sensing system needs to be modeled as a neural network layer, where some variables are set as trainable. For this layer modeling of the CASSI and MCFA system, let the observation model of the CSI fusion be

$$\mathbf{y}_c = \mathbf{D}_c\mathbf{P}\mathbf{T}_c\mathbf{f} + \omega_c$$

$$\mathbf{y}_m = \mathbf{T}_m\mathbf{D}_m\mathbf{f} + \omega_m,$$

the trainable variables are the entries of the coded aperture $\mathbf{T}_c$ in CASSI and the colored coded aperture $\mathbf{T}_m$ for MCFA. Then, defining the constraint for the training of these parameters to implementable values, for both systems, the entries must be in the set of $\{0,1\}$, recall that $\mathbf{T}_m$ and $\mathbf{T}_c$ are diagonalized matrices of the entries of $\mathbf{t}_m$ and $\mathbf{t}_c$ respectively, consequently, the feasible sets $\mathscr{R}_m$ and $\mathscr{R}_c$ can be defined as

$$\mathscr{R}_m = \left\{\mathbf{t}_m | \mathbf{t}_m \in \{0,1\}^{MNL_d}\right\}$$

$$\mathscr{R}_c = \left\{\mathbf{t}_c | \mathbf{t}_c \in \{0,1\}^{M_dN_d}\right\}.$$

To achieve a differential parametrization of this constraint, it was used a regularization

function on the loss function proposed in Bacca et al. (2021), which is given by:

$$R(\mathbf{x}) = \sum_i (1 - \mathbf{x}_i)^2 (\mathbf{x}_i)^2, \tag{23}$$

where the limits of the summation depend on the number of elements in $\mathbf{x}$. Note that the function $R(\mathbf{x})$ has roots in 0 and 1, then the function is minimized the values $\mathbf{x}_i$ trend to 0 or 1 as shown in Figure 9

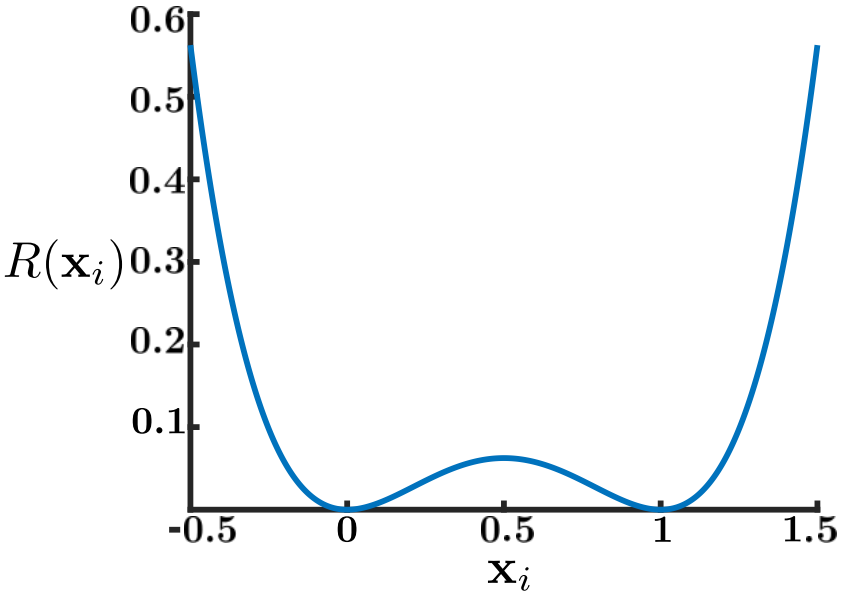

Figure 9. Differentiable regularization function for constraining the trainable parameters of the sensing layers to 0 or 1

4.2. Unrolling Fusion Network

In order to give a interpretability of the network M_θ in (22), the proposed unrolling approach for CSI fusion can be defined as follows. First, let define the optimization problem for the CSI

fusion

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}} ||\mathbf{H}_m\mathbf{f} - \mathbf{y}_m||_2^2 + ||\mathbf{H}_c\mathbf{f} - \mathbf{y}_c||_2^2 + \tau R(\mathbf{f}), \tag{24}$$

where the effect of the regularization term $R(\mathbf{f})$ is replaced by a spectral-spatial prior network as explained below. Introducing an auxiliary variable $\mathbf{h} \in \mathbb{R}^{MNL}$ and using the half quadratic splitting (HQS) method as in Wang et al. (2019a), equation (24) can be re-formulated as,

$$\{\hat{\mathbf{f}}, \hat{\mathbf{h}}\} = \arg\min_{\mathbf{f},\mathbf{h}} \frac{1}{2}||\mathbf{H}_m\mathbf{f} - \mathbf{y}_m||_2^2 + \frac{1}{2}||\mathbf{H}_c\mathbf{f} - \mathbf{y}_c||_2^2 + \tau R(\mathbf{h}) + \frac{\mu}{2}||\mathbf{f} - \mathbf{h}||_2^2, \tag{25}$$

where μ is a penalty parameter. The minimization of the two variables $\mathbf{f}$ and $\mathbf{h}$ can be found by solving the following sub-problems we have

$$\hat{\mathbf{f}}^{(k+1)} = \arg\min_{\mathbf{f}} \frac{1}{2}||\mathbf{H}_m\mathbf{f}^k - \mathbf{y}_m||_2^2 + \frac{1}{2}||\mathbf{H}_c\mathbf{f}^k - \mathbf{y}_c||_2^2 + \frac{\mu}{2}||\mathbf{f} - \mathbf{h}^k||_2^2, \tag{26}$$

$$\hat{\mathbf{h}}^{(k+1)} = \arg\min_{\mathbf{h}} \frac{\mu}{2}||\mathbf{f}^{k+1} - \mathbf{h}||_2^2 + \tau R(\mathbf{h}), \tag{27}$$

where $\mathbf{f}^k$ and $\hat{\mathbf{h}}^k$ denotes the estimation of $\mathbf{f}$ and $\mathbf{h}$ in the iteration k. First, the $\mathbf{h}$ sub-problem can be solved using a proximal operator of the image prior. For instance, if the is prior is the sparsity of the signal, the proximal operator can be a hard-thresholding operator Blumensath and Davies (2009)

or a soft-thresholding Beck and Teboulle (2009). Here, the prior is a deep prior which instead of learning it explicitly, it is only necessary to learn a proximal solver which is a sub-network in the unrolling network. Therefore, denote $S_{\theta^k}(\cdot)$ the deep proximal operator where θ^k are parameters of the CNN in the iteration k, consequently, the iterations of $\mathbf{h}$ are given by

$$\mathbf{h}^{(k+1)} = S_{\theta^{(k+1)}}(\mathbf{h}^k). \tag{28}$$

Then, the $\mathbf{f}$-problem can be solved using a gradient descent method, where each iteration is given by

$$\hat{\mathbf{f}}^{(k+1)} = \mathbf{f}^k - \lambda[\mathbf{H}_m^T(\mathbf{H}_m\mathbf{f}^k - \mathbf{y}_m) + \mathbf{H}_c^T(\mathbf{H}_c\mathbf{f}^k - \mathbf{y}_c) + \mu(\mathbf{f}^k - \mathbf{h}^k)], \tag{29}$$

with λ as the gradient descent step size. Finally, combining equation (28) and (29) the recursion of the algorithm is given by

$$\hat{\mathbf{f}}^{(k+1)} = \mathbf{f}^k - \lambda^k\left[\mathbf{H}_m^T(\mathbf{H}_m\mathbf{f}^k - \mathbf{y}_m) + \mathbf{H}_c^T(\mathbf{H}_c\mathbf{f}^k - \mathbf{y}_c) + \mu^k(\mathbf{f}^k - S_{\theta^k}(\mathbf{f}^k)\right]. \tag{30}$$

Note that the optimization parameters λ and μ also varies in each iteration of the algorithm. These parameters are also learned in the training of the network yielding in a E2E training of the reconstruction process. The recursion of (30) is shown in Figure 10. Finally, the initialization of

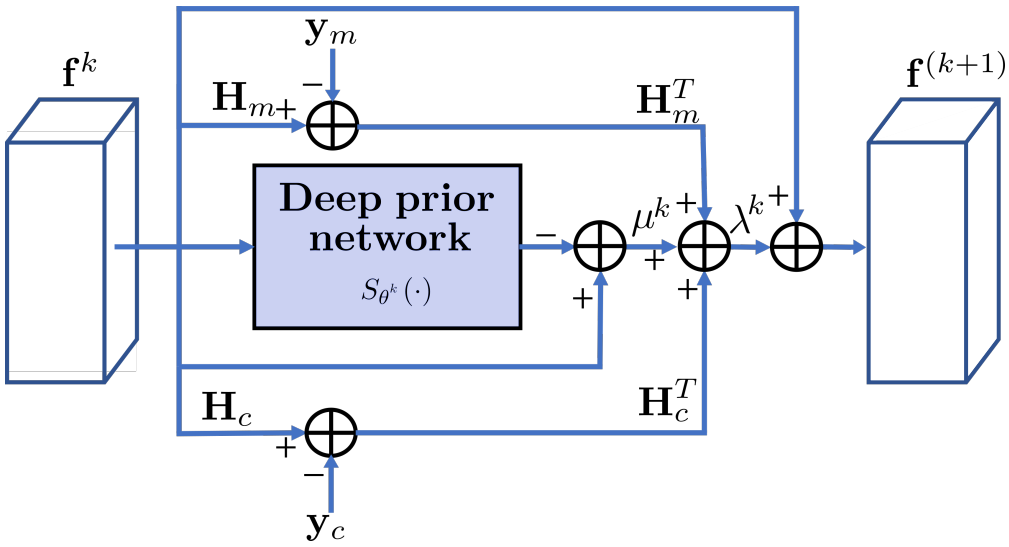

Figure 10. Recursion of the unrolling algorithm where in each stage is learned an optimal proximal operator S_{θ^k} and the parameters λ^k and μ^k

the unrolling network is given by:

$$\mathbf{f}^0 = \frac{1}{2}\left(\mathbf{H}_m^T\mathbf{y}_m + \mathbf{H}_c^T\mathbf{y}_c\right). \tag{31}$$

4.3. Deep spatial-spectral prior network

This section will detail the network that performs the proximal mapping in (28). . This network has two parts; one aims to exploit the spatial information, which is a kind of U-Net network Ronneberger et al. (2015) where has an encoder stage, where some features maps are extracted from the image until obtaining a latent space and a decoder which uses the features maps of the decoder stage by concatenating them into the channel dimension. The second part is a convolutional layer with dimension $1 \times 1 \times L$ which refines the spectral information. A graphical representation of the deep prior network is shown in Figure. 11. The following table summarizes the layers of the

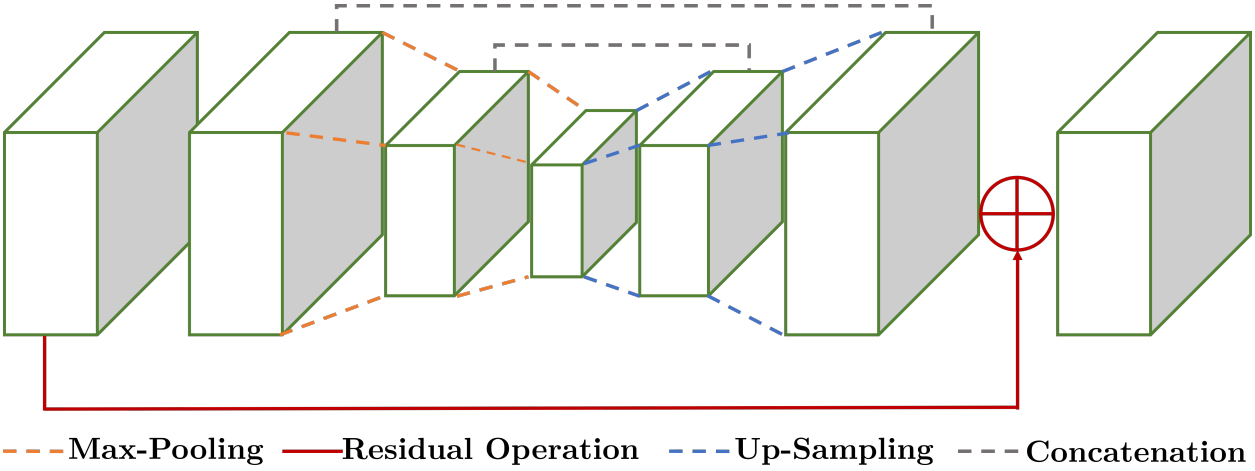

Figure 11. Deep prior network. Conformed by a encoder-decoder architecture for the spatial resolution and a spectral refinement layer.

deep spatial-spectral network

# Layer	Type	Description	Output Dimensions	Connections
1	2D Convolutional	Convolutional Layer with kernel dimensions (3,3,L)	$(M,N.L)$	Input
2	2D Convolutional	Convolutional Layer with kernel dimensions (3,3,L)	$(M,N.L)$	1
3	2D Max-Pooling	Max-Pooling with pool size (2,2)	$(\frac{M}{2},'\frac{N}{2},L)$	2
4	2D Convolutional	Convolutional Layer with kernel dimensions (3,3,L)	$(\frac{M}{2},\frac{N}{2},L)$	3
5	2D Max-Pooling	Max-Pooling with pool size (2,2)	$(\frac{M}{4},\frac{N}{4},L)$	4
6	2D Convolutional	Convolutional Layer with kernel dimensions (3,3,L)	$(\frac{M}{4},\frac{N}{4},L)$	5
7	2D Up-Sampling	Up-Sampling with size of (2,2)	$(\frac{M}{2},\frac{N}{2},L)$	6
8	Concatenation	Feature maps concatenation in the channel dimension	$(\frac{M}{2},\frac{N}{2},2L)$	7,3
9	2D Convolutional	Convolutional Layer with kernel dimensions (3,3,L)	$(\frac{M}{2},\frac{N}{2},L)$	8
10	2D Up-Sampling	Up-Sampling with size of (2,2)	$(M,N.L)$	9
11	Concatenation	Feature maps concatenation in the channel dimension	$(M,N,2L)$	10,1
12	2D Convolutional	Convolutional Layer with kernel dimensions (3,3,L)	$(M,N.L)$	11
13	Add	Residual Operation	$(M,N.L)$	11,Input
14	2D Convolutional	Convolutional Layer with kernel dimensions (1,1,L)	$(M,N.L)$	13

Table 1. Layers and its description of the deep spatial-spectral prior network

An overview of the proposed approach can be depicted in Figure. 8

4.4. Loss Function

An important factor in the deep learning approaches is the loss function since it determines how the trainables parameters should vary to improve the performance of the network. Depending of the network function one should choose. For instance if the network function is regression a

proper loss function is the mean squared error (MSE) or for classification a cross-entropy function should be used. This functions aims to compare the ground truth value with the estimated value and then propagate the error throughout the network. In this case, it was proposed a multiple loss for the reconstruction task. This loss is composed by

$$\mathscr{L} = \mathscr{L}_{spatial} + \mathscr{L}_{spectral}, \tag{32}$$

where the spectral component enforces a spectra fidelity and the spatial loss allows a visual enhancement of the estimated data cube. The first one is based on the spectral angle mapper (SAM) factor which is a valued parameter in spectral image application Yuhas et al. (1992); Zhuang et al. (2016). Denote $\mathbf{a}_{i,j}$ and $\mathbf{b}_{i,j}$ the spectral signatures of the spectral image $\mathbf{A}$ and $\mathbf{B}$ at the pixel (i, j), the spectral signatures can be seen as vectors of the dimension equals to the spectral bands number. The SAM metric of these spectral signatures is expressed as

$$SAM = \cos^{-1}\left(\frac{\mathbf{a}_{i,j} \cdot \mathbf{b}_{i,j}}{\|\mathbf{a}_{i,j}\|\|\mathbf{b}_{i,j}\|}\right), \tag{33}$$

where $\|\cdot\|$ is the vector norm and $\cdot$ refers to the dot product between vectors. It can be seem that the more simular are the vectors $\mathbf{a}_{i,j}$ and $\mathbf{b}_{i,j}$, the argument of the cosine fucntion tends to 1 and therefore the SAM value tends to 0. Consequently, the it is desired to achieve the following

condition

$$1 = \frac{\mathbf{a}_{i,j} \cdot \mathbf{b}_{i,j}}{\|\mathbf{a}_{i,j}\| \|\mathbf{b}_{i,j}\|}, \tag{34}$$

according with this, the spectral loss function can be define as following

$$\mathscr{L}_{spectral} = \sum_{i}^{M} \sum_{j}^{N} \|\hat{\mathbf{f}}_{i,j}\| \|\mathbf{f}_{i,j}\| - \hat{\mathbf{f}}_{i,j} \cdot \mathbf{f}_{i,j}, \tag{35}$$

where $\hat{\mathbf{f}}_{i,j}$ and $\mathbf{f}_{i,j}$ is the spectral signature at the pixel (i, j) of the estimated spectral image and the ground truth image respectively. For the spatial loss, it was used according to the main results of Zhao et al. (2017), where they proposed to use a combination of two losses for the spatial enhancement. They employ the Multi-Scale Structural Similarity Index Measure (MS-SSIM) which is an improved version of the Structural Similarity Index Measure (SSIM) Wang et al. (2004) which measures the perceived quality of an image, in this case, compared with a ground truth image. This metric takes into account structural information instead that absolute error metrics such as MSE or the Peak Signal to Noise Ratio (PSNR). The main improvement of the MS-SSIM concerning SSIM is that it introduces supplies more flexibility than single-scale methods in incorporating the variations of image resolution and viewing condition Wang et al. (2003). The other loss function that is used for the spatial enhancement is the ℓ_1 norm which encourages color preservation and luminance while the MS-SSIM high-frequency contrast Zhao et al. (2017). Therefore the spatial

loss is given by

$$\mathscr{L}_{spatial} = \frac{1}{MNL}\sum_{k}^{L}\sum_{i}^{M}\sum_{j}^{N}(1 - MS\text{-}SSIM(f_{i,j,k}, \hat{f}_{i,j,k})) + ||\mathbf{f} - \hat{\mathbf{f}}||_1. \tag{36}$$

Additionally, the regularization therm of the sensing layers (23) is added to the loss function. Then, the total loss of the network is given by

$$\begin{aligned}\mathscr{L} =& \rho_1 \frac{1}{MNL}\sum_{k}^{L}\sum_{i}^{M}\sum_{j}^{N}(1 - MS\text{-}SSIM(f_{i,j,k}, \hat{f}_{i,j,k})) + ||\mathbf{f} - \hat{\mathbf{f}}||_1 + \rho_2 \sum_{i}^{M}\sum_{j}^{N} \|\hat{\mathbf{f}}_{i,j}\|\|\mathbf{f}_{i,j}\| - \hat{\mathbf{f}}_{i,j} \cdot \mathbf{f}_{i,j} \\ &+ \rho_3 \left(\sum_{i}^{M_d}\sum_{j}^{N_d}(1 - \mathbf{t}_{c_{i,j}})^2(\mathbf{t}_{c_{i,j}})^2 + \sum_{k}^{L_d}\sum_{i}^{M}\sum_{j}^{N}(1 - \mathbf{t}_{m_{i,j,k}})^2(\mathbf{t}_{m_{i,j,k}})^2 \right),\end{aligned} \tag{37}$$

where ρ_1, ρ_2, ρ_3 are weighting hyper-parameters which can be selected using cross validation. These parameters weigh the how much will be affected the loss by each therm, for instance, if $\rho_3 >> \rho_1, \rho_2$ the training will prioritize the binarization of the coded apertures over the reconstruction quality or vice-verse.

Additionally, due to the depth of the final unrolling network, the network suffers of vanishing of the gradient in the back-propagation process yielding in a poor training of the parameters of the first stages resulting in a bad estimation of the data cube. For instance, the inception network in Szegedy et al. (2015) address this problem by computing the loss function in a intermediate layer of the network and back-propagate it from that layer. Similarly, here we propose to compute the loss function at the end of each stage, allowing a more accurate estimation in the firsts stages of the unrolling network, thus allowing that the algorithm converges with fewer stages.

5. Simulation Results

5.1. Datasets and pre-processing

For the training of the unrolling network, two well-known dataset were used; the ARAD dataset Arad et al. (2020)And the ICVL dataset Arad and Ben-Shahar (2016) which is composed by 201 hyperspectral images

5.1.1. ARAD dataset:. This was used in the NTIRE 2020 Challenge on Spectral Reconstruction from an RGB Image which contains 510 hyperspectral images in a spectral range of 400nm or 700nm with 31 spectral bands and 482×512 pixels. Figure. 12 shows 16 samples of the ARAD dataset where it is noticeable the diversity of the scenes in the dataset.

5.1.2. ICVL dataset:. This dataset contains 201 hyperspectral images collected at 1392×1300 spatial resolution over 519 spectral bands 400-1000nm. It has images of rural, urban, indoor scenes, among others Figure 13. To match the spectral range of the ARAD dataset, it was chosen the spectral bands in the visible range 400-700nm

Both datasets were spatially resized to 512×512 and it was used 20 spectral bands in the range of 450-650 nm, for the further experimental implementation. Also a normalization was used to keep all the values of the image in [0,1] for a better stability in the algorithm. Further, in the training process, different random patches of size 256×256 of the images were used in each training epoch. Therefore, in the model, the dimensions would be $M = N = 256$ and $L = 20$

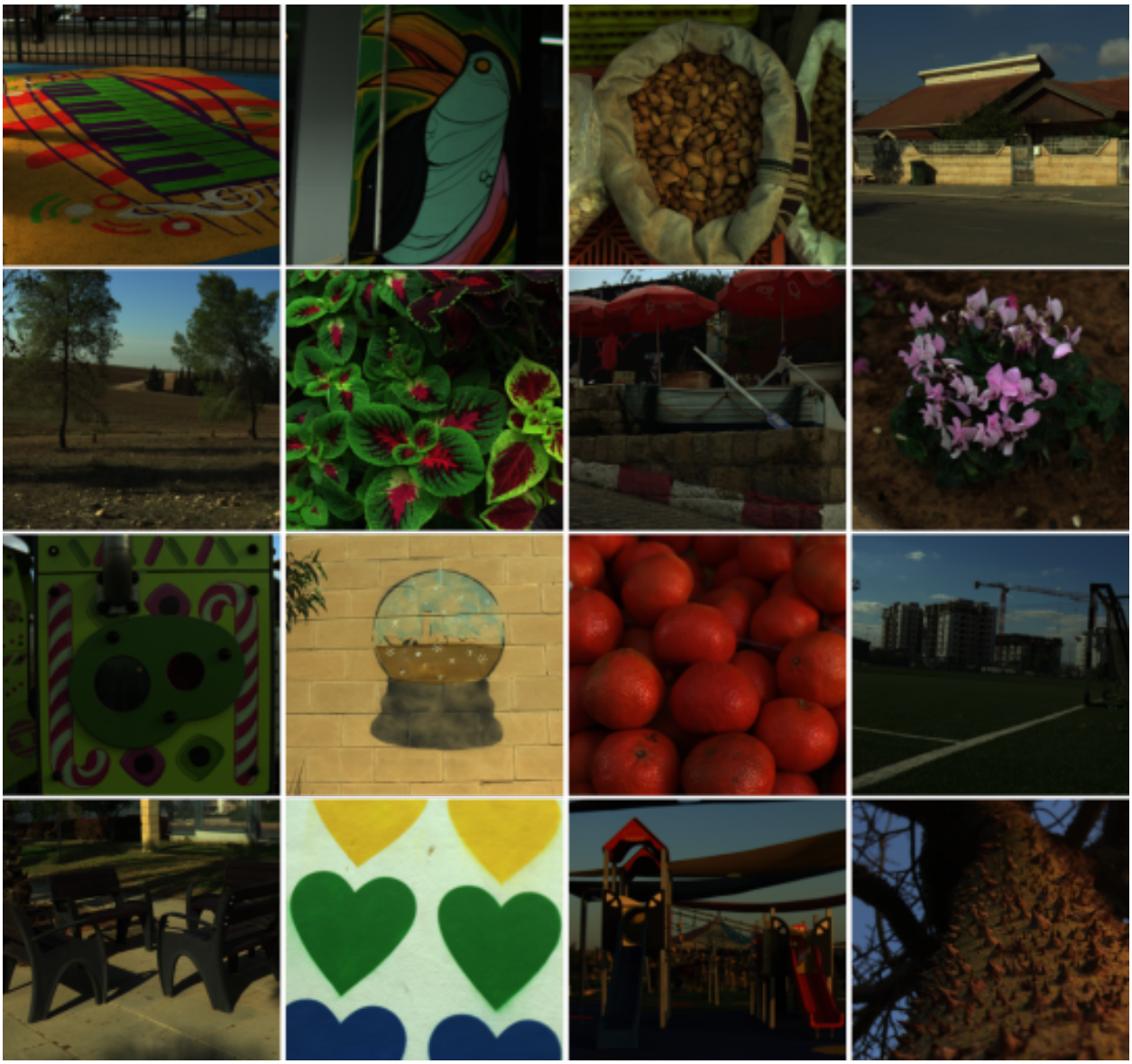

Figure 12. RGB representation of 16 samples of the ARAD dataset

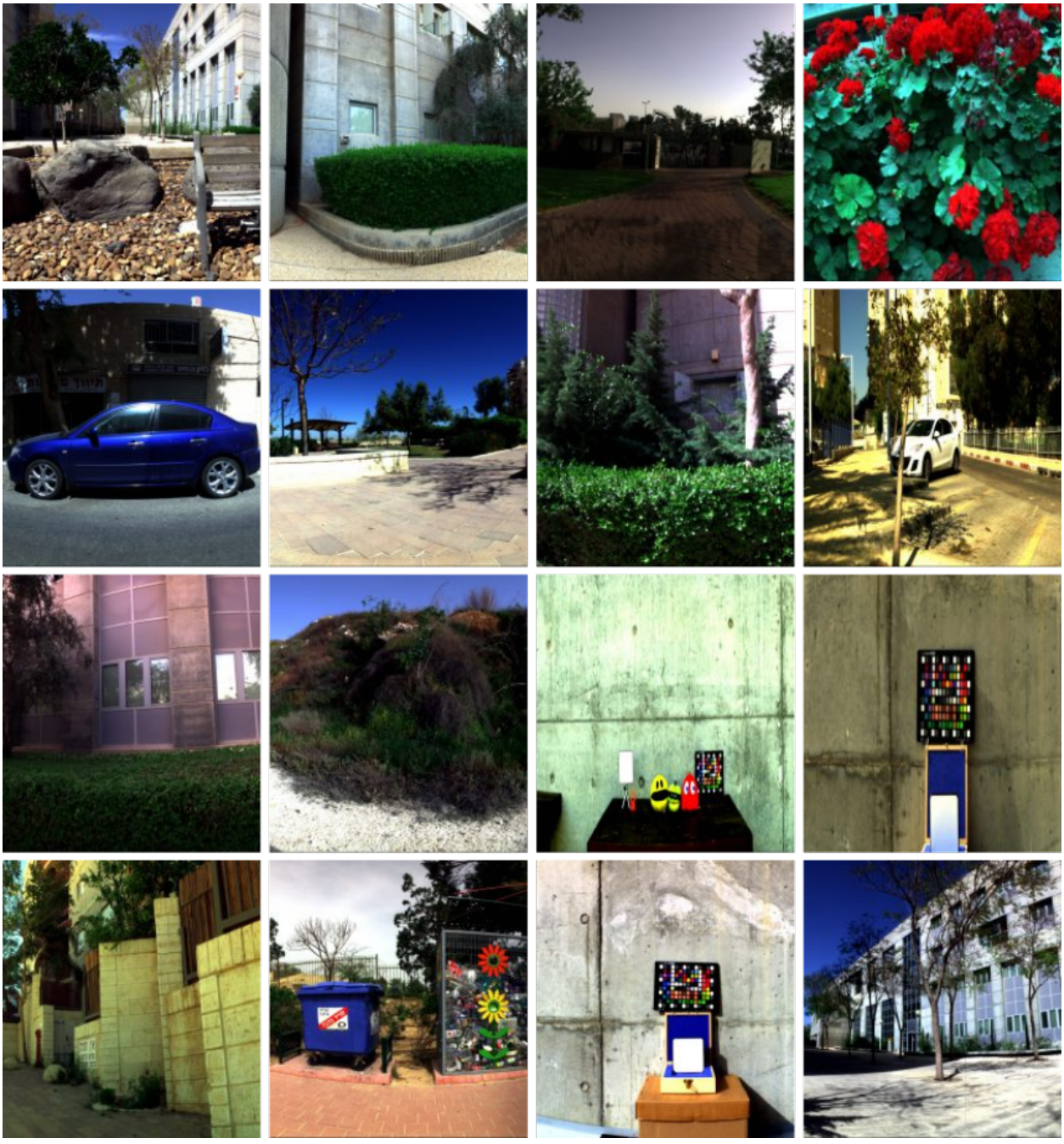

Figure 13. RGB representation of 16 samples of the ICVL dataset

5.2. Metrics:

To measure the reconstruction quality the following metrics were used the peak-signal-to-noise-ratio (PSNR) Horé and Ziou (2010), the structural similarity index measure (SSIM) Wang et al. (2004), dimensionless global relative error of synthesis (ERGAS) Renza et al. (2013), and the root mean square error (RMSE) Horé and Ziou (2010).

1. RMSE: Is a standard metrics to measure the difference between a predicted value and its corresponding ground truth. It is always positive and for values closer to 0, the better the predicted value. It is defined as

$$\text{RMSE}(\mathbf{f},\hat{\mathbf{f}}) = \sqrt{\frac{1}{MNL}\sum_{i}^{MNL} |\mathbf{f}_i - \hat{\mathbf{f}}_i|^2}, \tag{38}$$

   where $\mathbf{f}_i$ is the i-th pixel of the image

2. PSNR: Measured in dB, is defined as the logarithm of the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation so that a higher value indicates superior quality of fusion Gelvez and Arguello (2020). And is expressed as

$$\text{PSNR}(\mathbf{f},\hat{\mathbf{f}}) = 10\log_{10}\left(\frac{\max(\mathbf{f})^2}{\text{RMSE}(\mathbf{f},\hat{\mathbf{f}})^2}\right). \tag{39}$$

3. ERGAS: normalized average error of each band of processed image. Lower value of ERGAS indicates that the estimated image is similar to the reference image Jagalingam and Hegde

(2015). The ERGAS metric is given by

$$\text{ERGAS}(\mathbf{f},\hat{\mathbf{f}}) = \frac{100}{r_s r_\lambda}\sqrt{\sum_i^L \frac{\text{RMSE}(\mathbf{f}_i,\hat{\mathbf{f}}_i)^2}{\mu_i^2}}, \tag{40}$$

where $\mathbf{f}_i$ denote the i-th spectral band of $\mathbf{f}$ and the parameters r_s and r_λ are the spatial and spectral decimation factors of both systems.

4. SSIM: As mentioned before, this metric measure the quality of the estimated image in terms of the degradation of the structural information instead of absolute errors. It is implemented on various windows of the image, denote $\mathbf{f}_X$ and $\hat{\mathbf{f}}_Y$ a window of $S \times S$ of the ground truth image and the estimated image respectively, then, the SSIM metric can be defined as

$$\text{SSIM}(\mathbf{f}_X,\hat{\mathbf{f}}_Y) = \frac{(2\mu_X\mu_Y + c_1)(2\sigma_{XY} + c_2)}{(\mu_X^2 + \mu_Y^2 + c_1)(\sigma_X^2 + \sigma_Y^2 + c_2)}, \tag{41}$$

where μ, σ are the mean value and variance of the window, σ_{XY} are the covariance of $\mathbf{f}_X$ and $\hat{\mathbf{f}}_Y$, $c_1 = (k_1 L)^2, c_2 = (k_2 L)^2$ are two variables to stabilize the division with weak denominator, L is the number of quantization levels of the image, k_1 and k_2 are hyperparameters, usually 0.01 and 0.03 respectively. For SSIM values close to 1 the quality of the estimated image is better.

5.3. Comparison methods

As comparison for state-of-the-art methods of compressive spectral image fusion, it was used the work in Vargas et al. (2018) which is a method based on convex optimization using

sparsity and total variation regularizers, and will be denoted as Sparse-Based Fusion (SBF). Also, some self-comparison was performed. One using a classic loss function as the mean-squared-error denoted proposed$_{MSE}$, other using just the loss in the last stage denoted proposed$_{SL}$, to show the improvement by learning the optimal coded apertures, a experiment was performed keeping fixed the coded apertures, this is denoted proposed$_{NT}$ and finally, the entire proposed method will be denoted as proposed$_{T}$.

5.4. Simulation Configuration

The Adam optimizer Kingma and Ba (2014) was used for the training of the unrolling network, the coded apertures were initialized with a normal random distribution, the regularization parameter to achieve binary values was set in $\rho_3 = 5$, the gradient step parameter was initialized $\lambda = 0{,}001$ and the penalty parameter $\mu = 0{,}01$. The learning rate of the network was initialized in 0.001 and it was halved every 40 epochs. The regularization parameters of the loss function were set $\rho_1 = 2, \rho_2 = 3$. For the experiments with fixed coded apertures, these were set following a Bernoulli distribution with $p = 0{,}5$. The following sections will show the simulation results for the datasets aforementioned. And finally, 200 epochs were employed for the training. It was used $K = 9$ stages in the unrolling network.

5.5. Simulation results for the ICVL dataset

From the 201 images of the dataset, 31 images with similar content were removed, the 180 remaining images were split 154 for training, 10 for validation and 16 for test. The code for the SBF method was provided by the authors, and the results presented are the mean values of the reconstruction of the training dataset. The Table. 2 shows the quantitative results for the mentio-

ned methods over the training set of the ICVL dataset. It can be observed that the entire proposed method outperformed the others methods. There is an exception, with the PSNR of the proposed$_{SL}$ method, which outperforms the proposed$_{T}$ method for very little value. The main improvement of multi-loss in the proposed$_{T}$ will be shown in the reconstruction quality along the stages in the proposed methods Figure 14. It can be seen that, the proposed methods which use the multiple loss strategy reaches the maximum value o close to it with fewer stages than the proposed method with single loss. For instace, comparing the proposed$_{SL}$ and proposed$_{T}$ which have a final similar reconstruction quality value, in the stage 6 the proposed$_{T}$ reaches a PSNR 40 dB while the proposed$_{SL}$ only have a PSNR 33 dB.

Method	**PSNR**	**SSIM**	**ERGAS**	**SAM**	**RMSE**
SBF	30.94	0.87	13.78	0.089	0.028
Proposed$_{MSE}$	38.31	0.965	10.56	0.153	0.023
Proposed$_{NT}$	39.76	0.978	6.402	0.451	0.017
Proposed$_{SL}$	**41.017**	0.977	10.17	0.068	0.015
Proposed$_{T}$	41.012	**0.981**	**6.179**	**0.041**	**0.014**

Table 2. Quantitative results of the test data reconstruction quality for the ICVL dataset

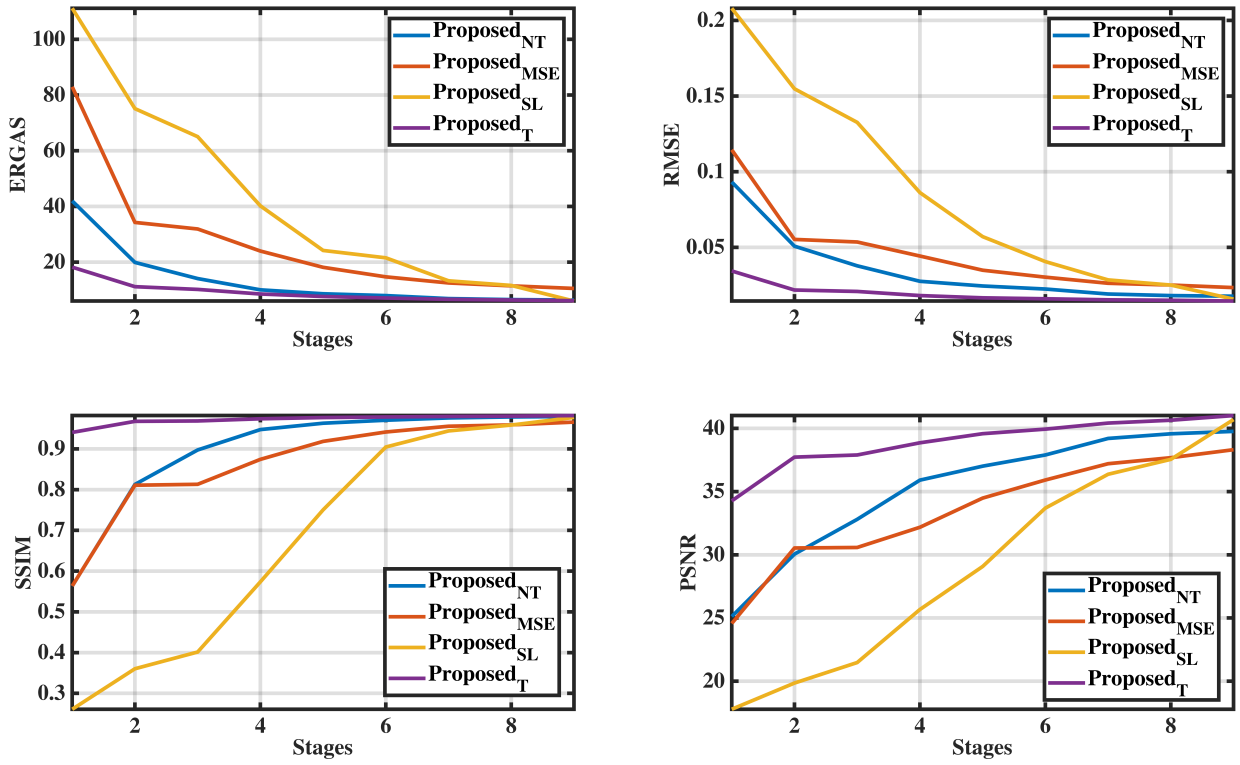

Figure 14. Metrics of the reconstruction quality along the stages of the proposed unrolling network. The superiority in the convergence of the algorithm is notorious in the proposed$_T$ for the ICVL dataset

5.6. Simulation results for the ARAD dataset

Here, 40 images with similar content were removed, thus, the dataset was split in 450 images for training, 5 images for validation and 15 images for testing. The Table 3 summarizes the results obtained with the ARAD dataset. In Figure 15 are shown the results along the stages. Here is also notorious the improvement in the convergence of the algorithm using the multiple loss compared with using a single loss at the end of the network.

Method	PSNR	SSIM	ERGAS	SAM	RMSE
SBF	31.943	0.846	15.393	0.098	0.033
Proposed$_{MSE}$	40.521	0.9704	10.921	0.068	0.022
Proposed$_{NT}$	33.563	0.8766	22.4408	0.215	0.0458
Proposed$_{SL}$	40.752	0.9752	9.452	0.053	0.018
Proposed$_{T}$	**41.5311**	**0.980**	**7.2482**	**0.032**	**0.014**

Table 3. Quantitative results of the test dataset reconstruction quality for the ARAD dataset

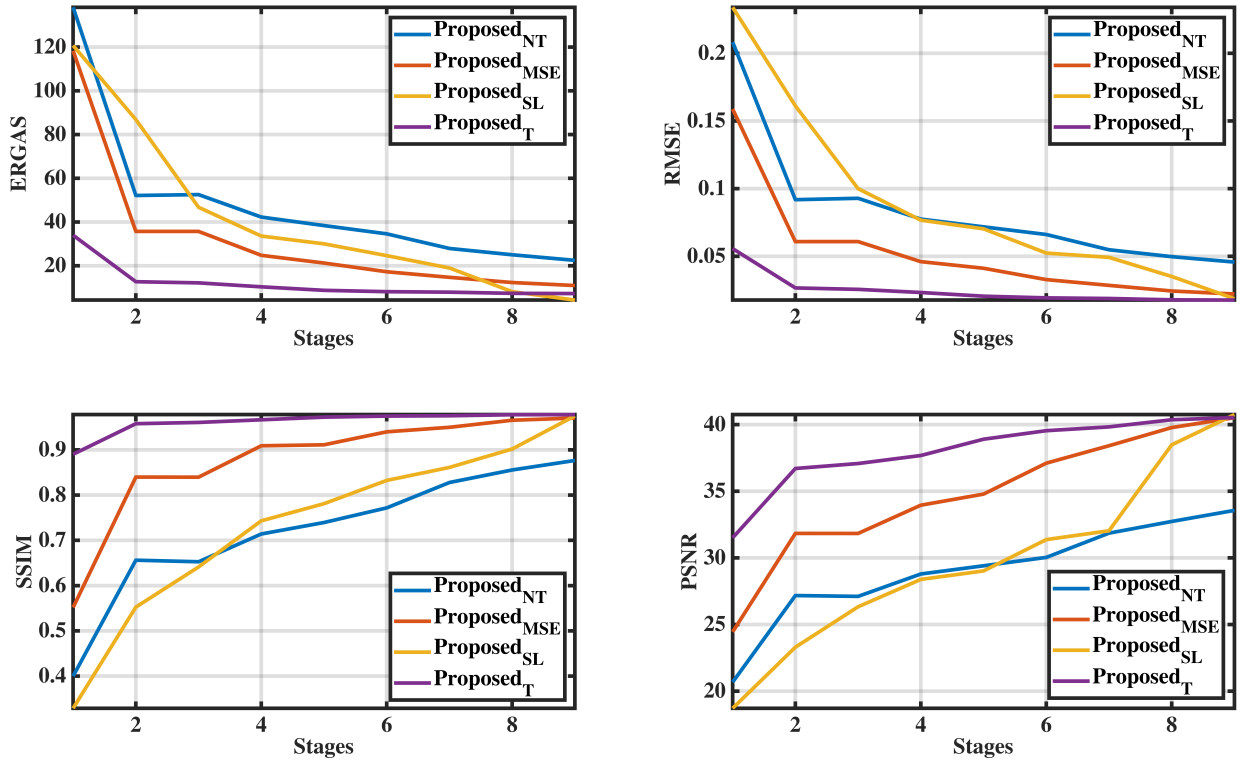

Figure 15. Metrics of the reconstruction quality along the stages of the proposed unrolling network. The superiority in the convergence of the algorithm is notorious in the proposed$_T$

5.7. Visual Results and Spectra Reconstruction

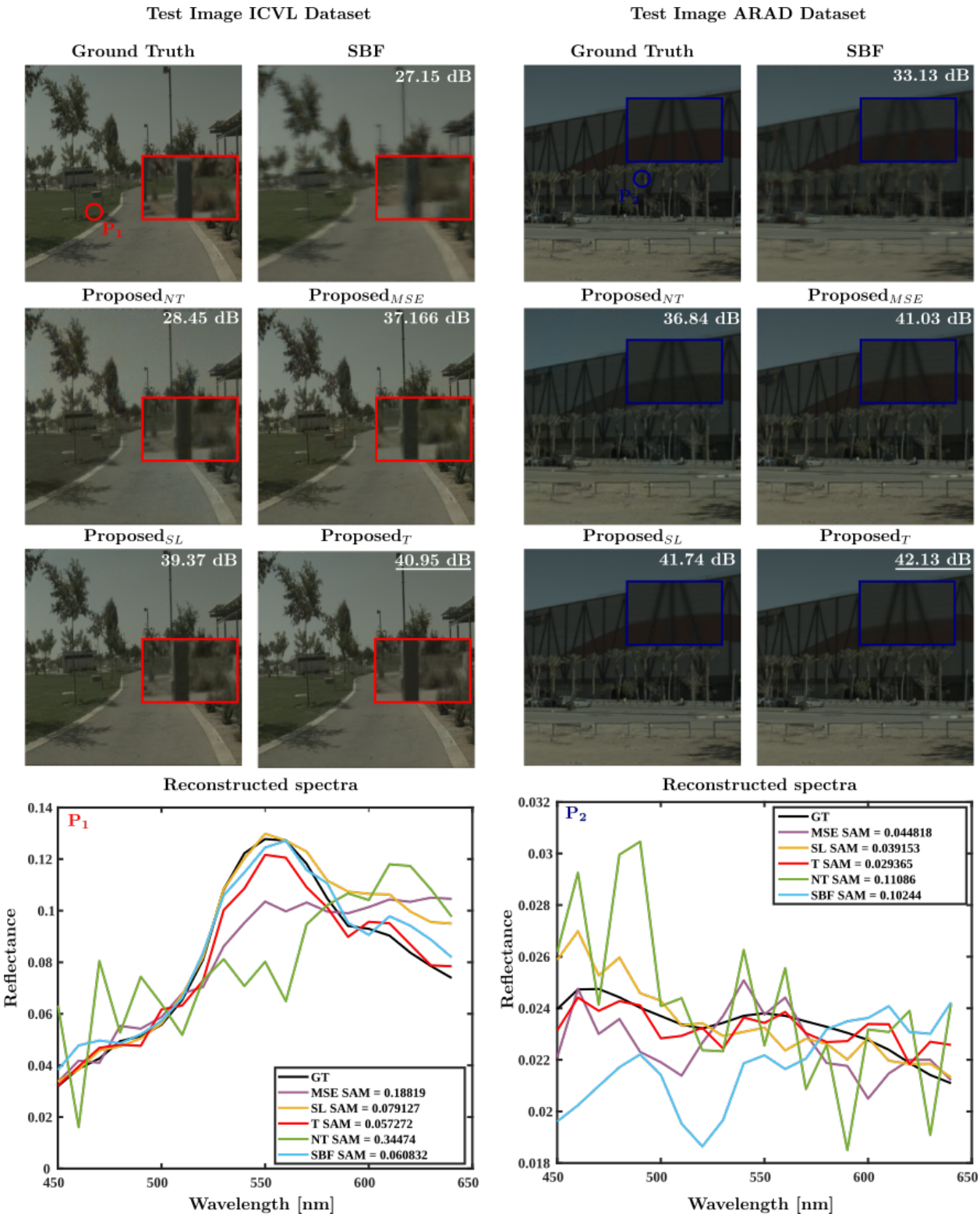

Figure 16. False RGB visualization of a test image of both datasets and the reconstructed spectra of a representative point in each image

For visual representation of a reconstructed image of both datasets with the mentioned methods, a false RBG representation was used through the spectral signatures of the respective colors red, green and blue. Also a gamma correction was used to improve the contrast in the visualization. A region of the image is zoomed to appreciate small details of the reconstructed image. Also the PSNR metric of the reconstructed image is displayed for a quantitative measure of the image quality respect the ground truth. Finally, a spectral signature of a representative point of the image, the grass of ICVL image and the orange of the building in the ARAD image. These results also shows that the entire proposed method (proposed$_T$) outperformed the other methods, with the lowest SAM values and more pleasant visual representation of the reconstructed images in both images.

5.8. Trained coded apertures and color filter array

In Figure 17 is shown the designed coded aperture with the end-to-end approach.

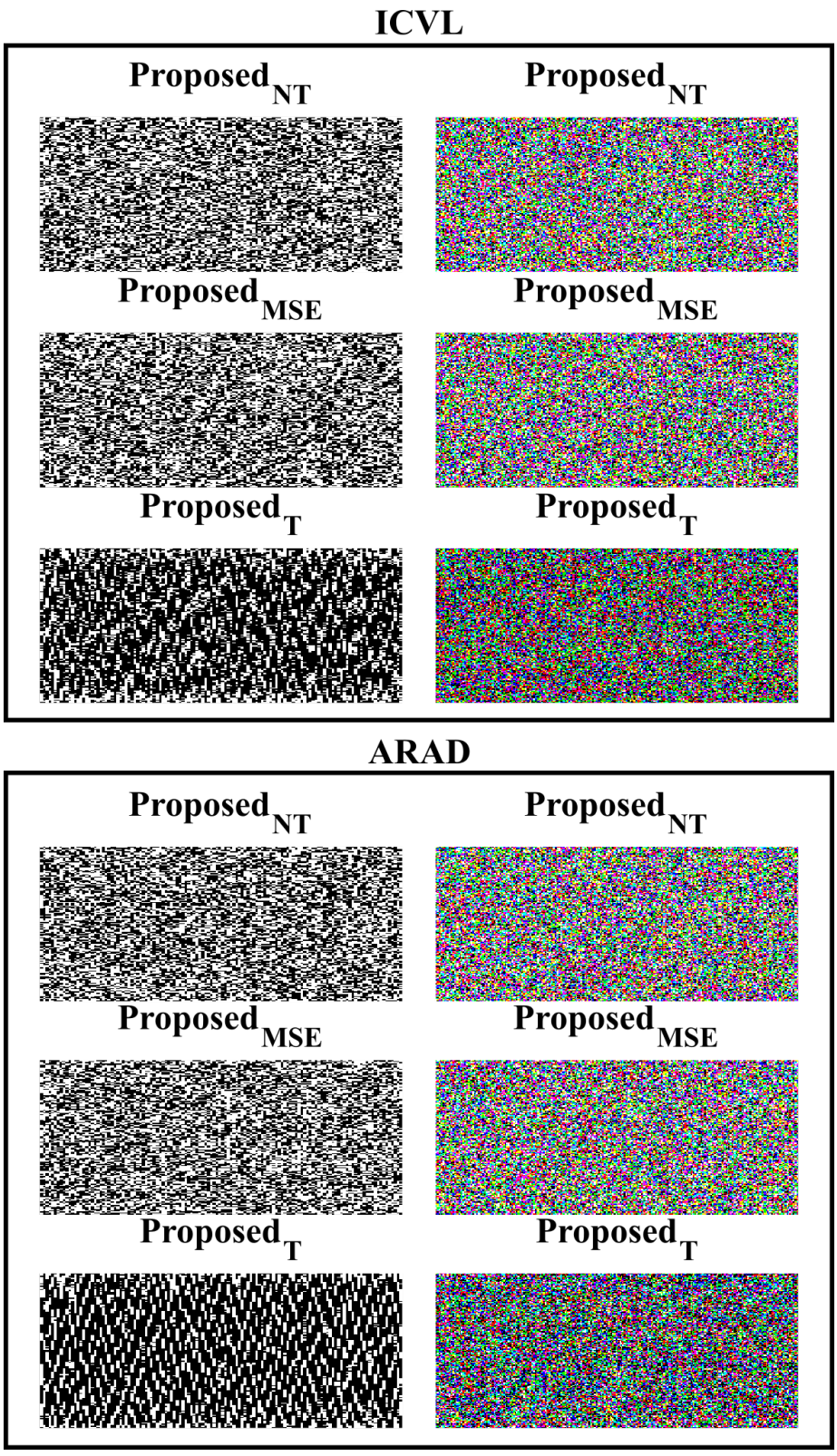

Figure 17. Trained coded apertures (left) and color filter arrays (right) with different trained models

6. Experimental Set-Up and Real Data Validation

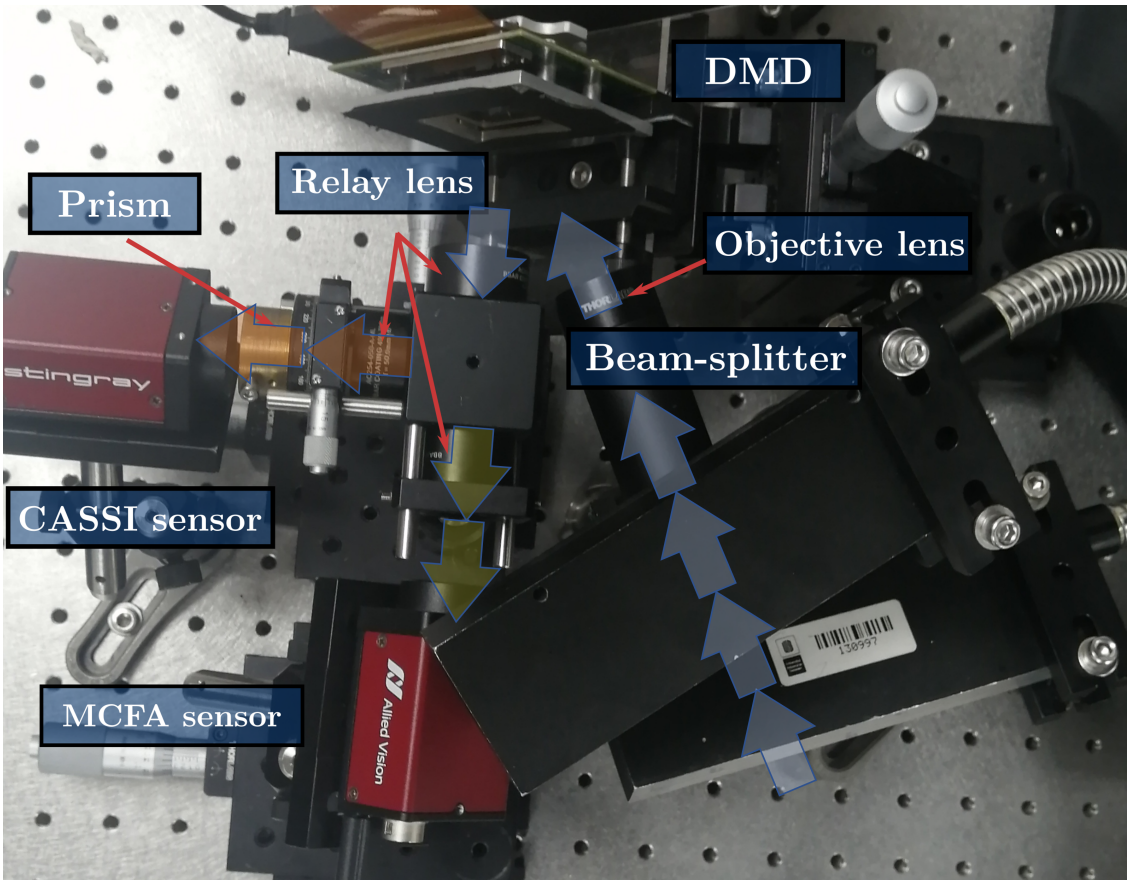

Figure 18. Optical architecture of the experimental prototype of the dual CSI system

The experimental prototype of the proposed dual CSI system was assembled in the optics laboratory of the High Dimensional Signal Processing (HDSP) research group

6.1. Optical elements

The following table summarizes the list of elements employed for the experimental prototype. Here some remarks of the configuration of the elements:

Item	Quantity	Description	Reference
Grayscale camera	2	2D optical sensors, to capture the projected measurements of both systems	F-145 Stingray
Digital Micromirror Device (DMD)	1	Programmable light modulator to create the coded apertures	Texas instruments DLP4130
Monochromator	1	Programmable light source at a determined wavelength	Cornerstone 130 1/8m.
Relay lens	2	Achromatic lens with focal length of 30mm and 50mm	Achromatic Lens LSB08 Series Thorlabs
Objective lens	1	Aperture adjustable lens with 8mm focal length	
Beam splitter	1	Optical element which divide the incident light source in two	Non-Polarizing Beams splitter CCM1-BS013 Thorlabs
Prism	1	Dispersive element which decompose the incident light into its spectral components	AMICI Prism

Table 4. List of the optical elements employed in the experimental prototype

- The control of the programmable elements (cameras, DMD, and monochromator) was made in MATLAB. Particularly, the cameras were controlled using the image acquisition toolbox, the monochromator was programmed using serial communication and the DMD was handled using a specialized API library for the device.

- The monochromator is programmed to illuminate at a determined wavelength. Nevertheless, the bandwidth (how much spectrum) of the output illumination depends on a pair of slits placed in the monochromator see Figure. 20. These slits limit the bandwidth of the light but reduce its intensity thus requiring to increase in the exposure time of the camera to counter this illumination issue and consequently increasing the acquisition time. It was found that using only one slit of 600 of [μm] brings the optimal relation between exposure time and intensity of the illumination.

- The DMD is a programmable array of micromirrors. Each micromirror has two states 0 or

1. This state just describes the angle in which the micromirror is set to see Figure 19. for an illustration of the DMD operation.

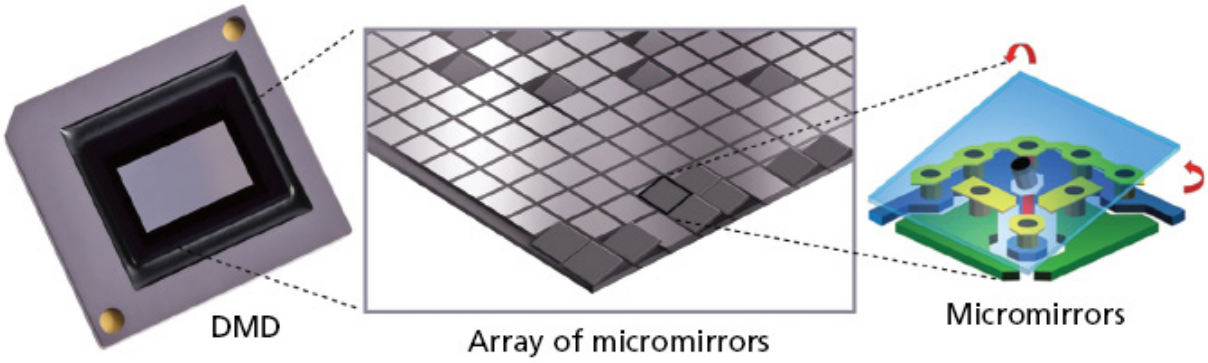

Figure 19. Illustration of the DMD operation. The 0 state sets the angle of the micromirror in which that part of the scene is blocked to the sensor, and the 1 state reflects the light into the sensor

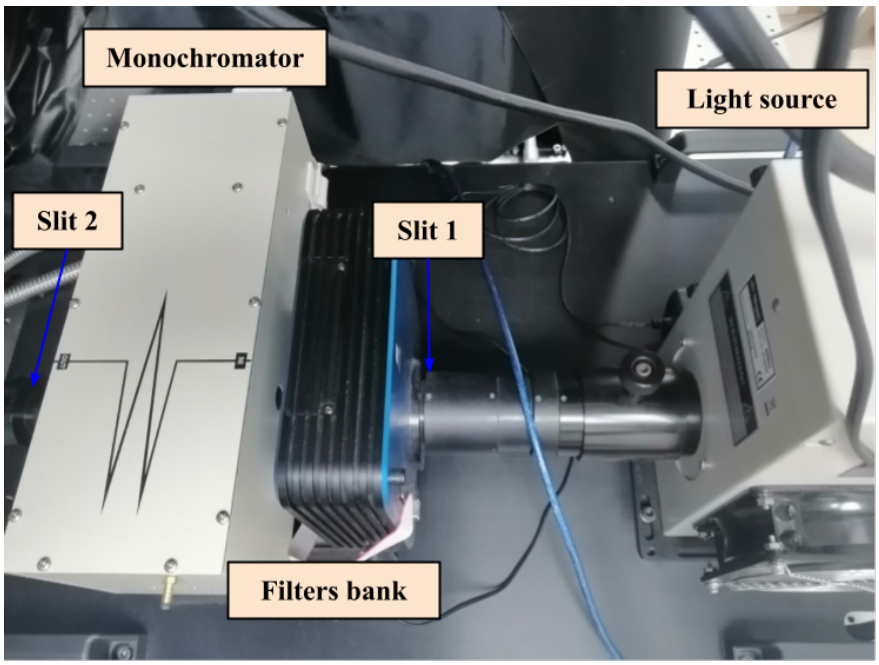

Figure 20. Structure of the monochromator. It has a light source, a bank of optical filters which are adjusted according to the wavelength range which is going to be used, the monochromator itself which contains a set of diffractive grating which decompose the white light of the source and selects the desired wavelength. The output light is transported through an optic fiber to the scene. Two slits limit the bandwidth of the emitted light

- Three parameters are adjusted in the cameras. The gain, shutter, and exposure time. After, some tests, the optimal values of these parameters were; gain = 100, shutter = 4095 and exposure time = 0.8 [s] for the MCFA sensor and 1 [s] for the CASSI sensor. This difference is due to the prism reduces the intensity of the incident light on the CASI sensor.

6.2. Assembly of the dual-arm system

The assembly of the optical architecture was carried out as follows

1. The first element that was placed in the DMD and then it was calculated the optimal lens and the respective distance of the object and the sensor. To clarify, in this stage, the DMD was the target. To this purpose, it was used the thin lens approximation equation which holds

$$\frac{1}{d_i} + \frac{1}{d_o} = \frac{1}{f}, \tag{42}$$

where d_i is the distance where the image is formed, therefore we need to place the sensor in that distance and d_o is the distance where the object should be placed.

2. After, it was placed the first sensor, the beam splitter was located in the space between the lens and the already set sensor and the second sensor was placed considering that the focal length remains but the optical path changes due to the beam splitter.

3. Then, with the 2 sensors correctly placed, the prism was located between the beam splitter and one of the sensors.

4. Then, the other side of the architecture is the one in charge of focusing the scene onto the

DMD, therefore, an objective lens of 8mm was used and due that this lens forms the image at a very close distance, a relay lens of 30mm was employed which is placed at a distance 2f form where the image is from an also 2f of the DMD, therefore, this lens copies the image formed by the objective lens into the DMD.

6.3. Calibration process

The sensors, DMD, and prism were rotated 45°. As mentioned before, the DMD works by inclining its micromirror to block or reflect the light this angle produces that the first stage of the system has to be aligned with this angle, the Figure. 21. To locate accurately the location of the sensors and the DMD, it was used a set of precision sliders.

Since the sensors have a pitch size of 6.5 [μm] and the DMD has a pitch size of 13.5 [μm], therefore one pixel of the DMD equals a 2×2-pixel area of the sensor. As mentioned before, the high spatial resolution dimensions are $M = N = 256$, we set the DMD with the following considerations

- **MCFA:** As this systems senses the spectral image with high spatial resolution, the DMD was used with a $M \times N$ coded aperture.

- **CASSI:** Considering that this system has a lower spatial dimension $M_d = N_d = 128$. Nevertheless, the DMD resolution was kept fixed in 256×256 for both systems. Then, to obtain a 128×128 coded aperture it was employed a Kronecker product between the low resolution coded aperture and a 2×2 matrix full of ones to replicate the 128×128 coded aperture into a 256×256. The Kronecker of the coded aperture $M_d \times N_d$ matrix $\mathbf{A}$ and the 2×2 matrix

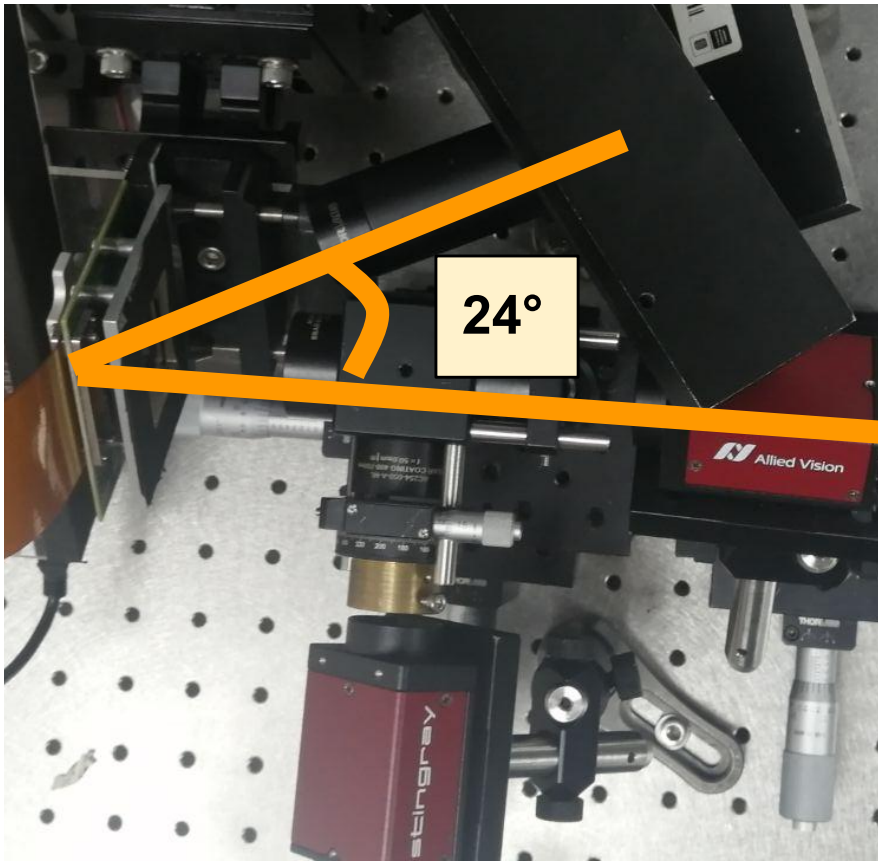

Figure 21. Alignment of the optical architecture with the DMD inclination angle

full of ones $\mathbf{B}$ results in $M \times N$ matrix given by:

$$A \otimes B = \begin{bmatrix} a_{1,1}\mathbf{B} & \cdots & a_{1,N_d}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{M_d,1}\mathbf{B} & \cdots & a_{M_d,N_d}\mathbf{B} \end{bmatrix} \tag{43}$$

therefore, each pixel of the coded aperture is replicated in 2×2 pixel area in the DMD.

- While the resolution of the DMD was set to 256×256, to keep the ratio of the pitch size

between the DMD and the sensor, the optical systems were calibrated to obtain a 512×512 image in the camera, and then a downsampling operation was used to convert the measurements into the respective dimensions for both systems (256×256 for MCFA and 128×147 for CASSI).

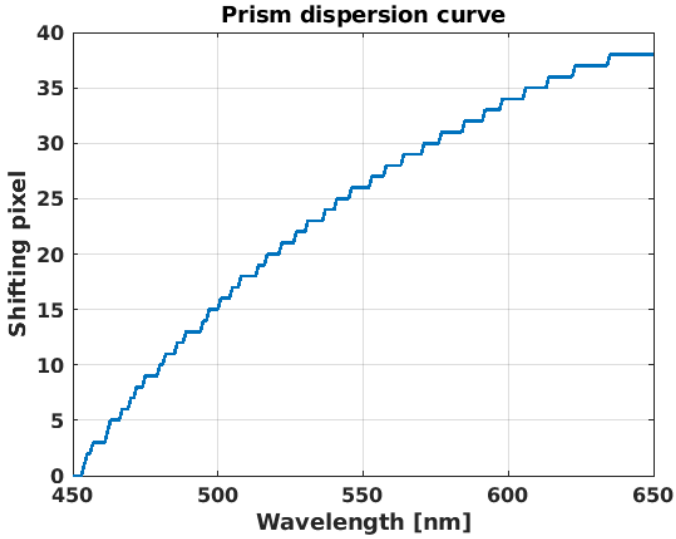

Figure 22. Characterization curve of the prism

Then, it was characterized by the prism dispersion curve. The curve discretizes the dispersion of the prism with the pixels of the sensor, therefore allowing to determine how many bands are going to be taken. Figure 22 shows the dispersion curve of the prism. This curve was obtained using as coded aperture a line and acquiring a withe scene with a wavelength sweep. Every change in the dispersion curve establishes a new spectral band, however, it needs to take into account that there is spatial downsampling of $r_s = 2$ in the system, therefore, the number of bands is determined for every two changes in the dispersion curve. Also, by only using the spectral range between 450 and 650 [nm], we obtained 20 spectral bands for the CASSI system and 10 spectral bands for the

MCFA system.

6.4. Scene capturing

The measurements of both systems were acquired using a wavelength sweep with the monochromator and sensing the coded image for each wavelength. The following subsections will be detailed the acquisition process of each system

6.4.1. MCFA. Here, the color filter array was emulated using the DMD, where, each wavelength is associated with a coded aperture. For this purpose, it was taken L_d shots of the scene, where in each shot the scene is illuminated with a different wavelength and the coded aperture corresponding to each wavelength is set to acquire the coded projection. The illumination of the determined wavelength can be seen as spectral filtering. Then, as a post-processing, it is concatenated each acquired shot and subsequently, the sum over the spectral dimension to obtain the compressive measurements see Figure. 23. Note that each pixel value on the coded aperture represents the spectral response of the color filter.

6.4.2. CASSI. Similar to the MCFA acquisition process, the CASSI sensing process it was uses a wavelength sweep of the scene. The main difference is that the coded aperture is the same for all wavelengths as this is just a block-unblock coding element. In the sensor every frame acquired is shifted according to the dispersion curve of the prism, and finally, all shots taken are sum in the spectral dimension.

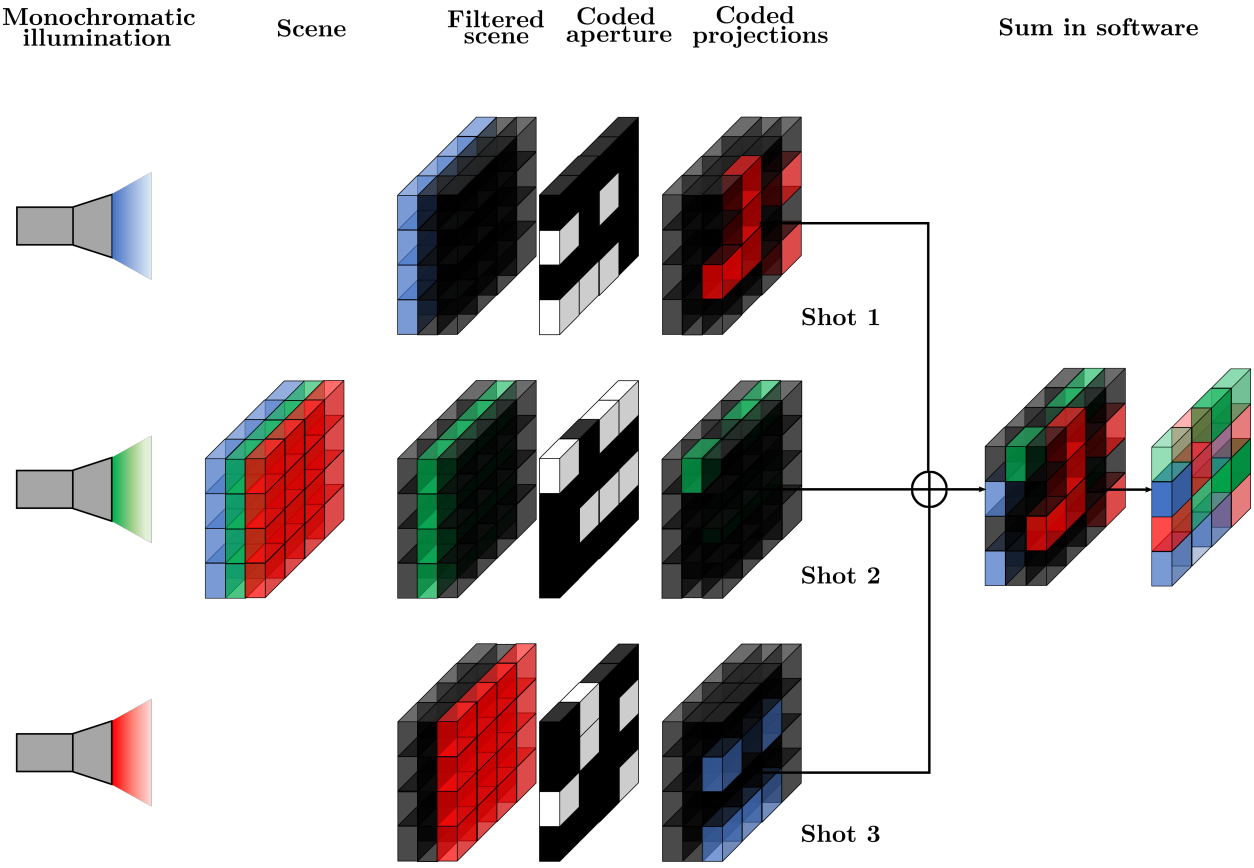

Figure 23. Scheme of the MCFA acquisition process for 3 spectral bands (red, green, and blue), wherein each shot the scene is illuminated with a determined monochromatic light and a coded aperture associated with the wavelength of the light source is used to codify the scene, then, all the measurements are concatenated and sum in the spectral dimension.

6.5. Post-processing

Two trained models were impremented, the ICVL Proposed$_{\text{NT}}$ and Proposed$_{\text{T}}$ to compared the results obtained with non-designed CA and designed CA. The calibrated sensing matrices are shown in Figure. 25.

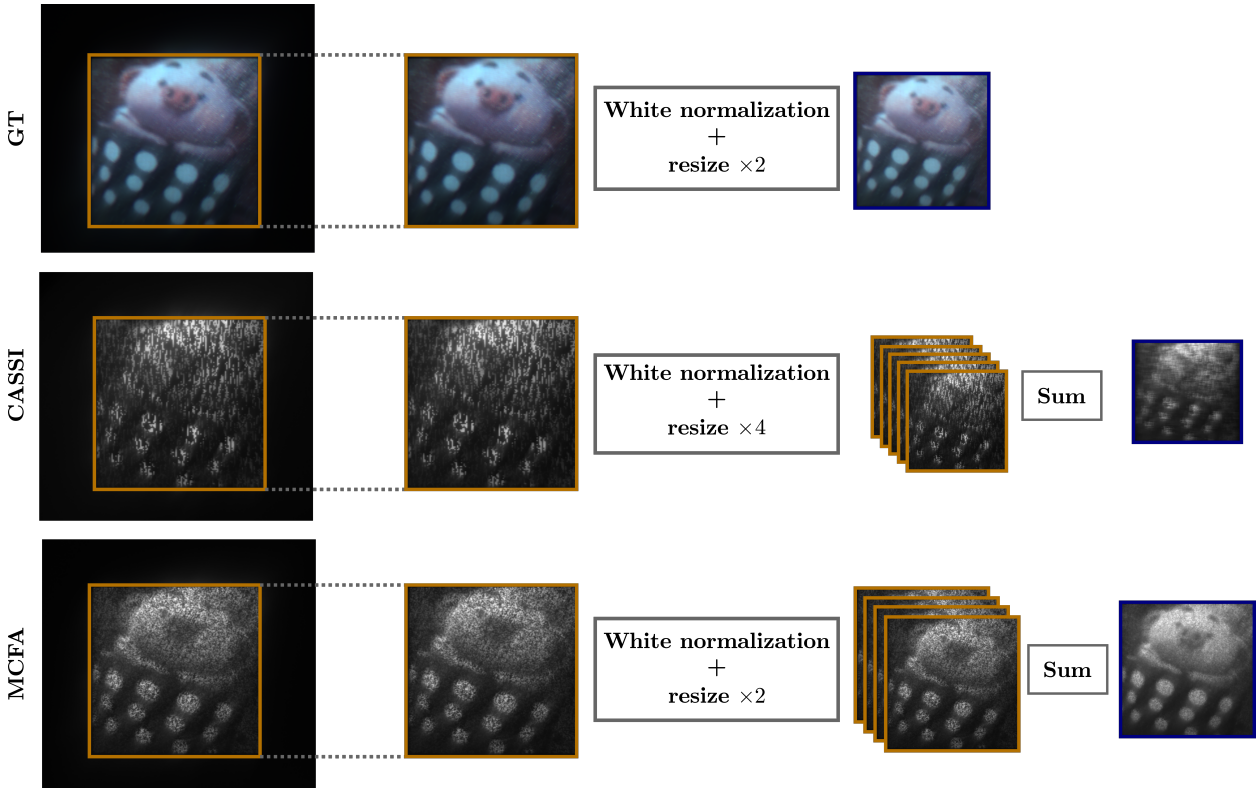

Figure 24. Post processing of the raw acquired data. The region of interest is cropped, then a normalizing by a white spectral signature for each sensor is used, then the coded measurements are sum in the third dimension to obtain the final compressive measurements

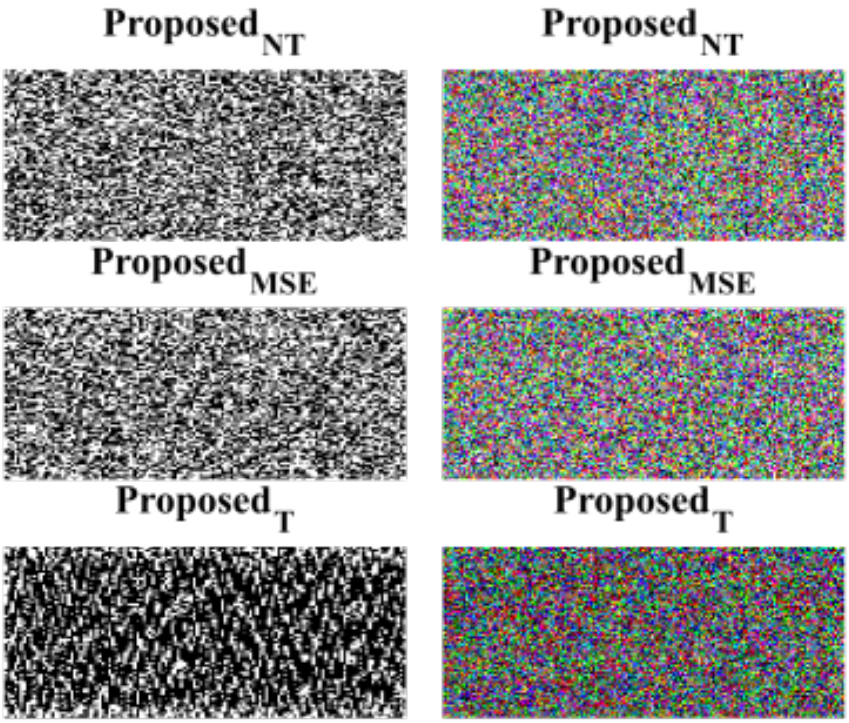

Figure 25. Calibrated Non-designed CA (random distribution) and designed obtained with a E2E approach for the CA-CASSI and CA-CFA, respectively.

It was captured three-set of measurements from 4 scenes; The MCFA and CASSI coded measurements and the ground truth of the scenes. The post-processing of the acquired images is described in Figure. 24.

6.6. Inference of the proposed reconstruction process

For real data inference, a re-training of the network was performed using four captured scenes, with data augmentation such as rotation and changing the contrast.

6.7. Results with captured data

To compare the proposed unrolled network, the alternating direction method of multiplier (ADMM) with its variant plug and play (PnP) is implemented Chan et al. (2016). This algorithm takes advantage of the ADMM structure by replacing one of its optimization steps with a denoising algorithm; this work uses a recursive filter denoiser Gastal and Oliveira (2011).

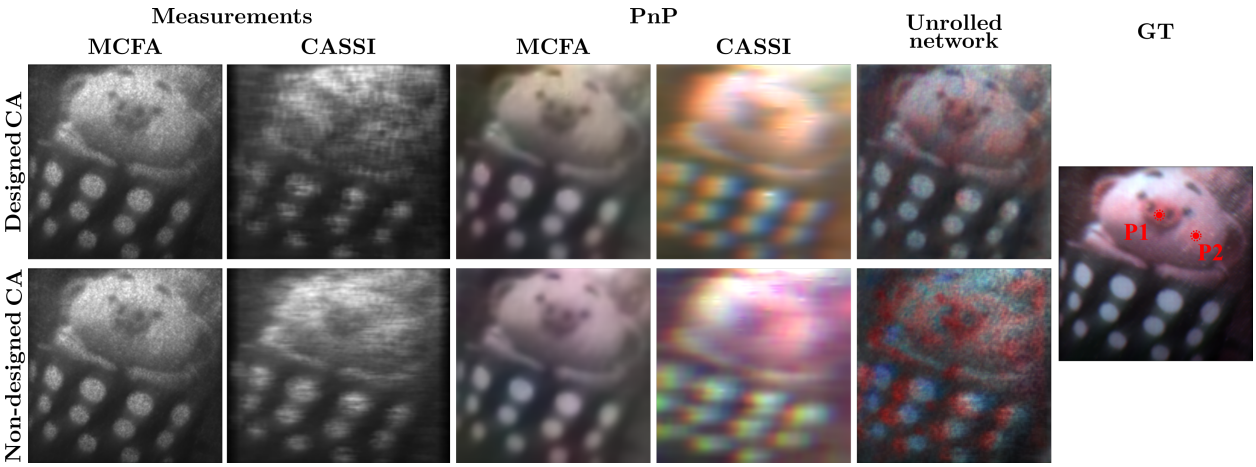

Figure 26. Visual representation of the compressive measurements and the reconstructed images using the PnP algorithm for the individual measurements and the proposed unrolled fusion network. (Top) shows the results obtained with the designed CAs and (bottom) with non-desing CAs

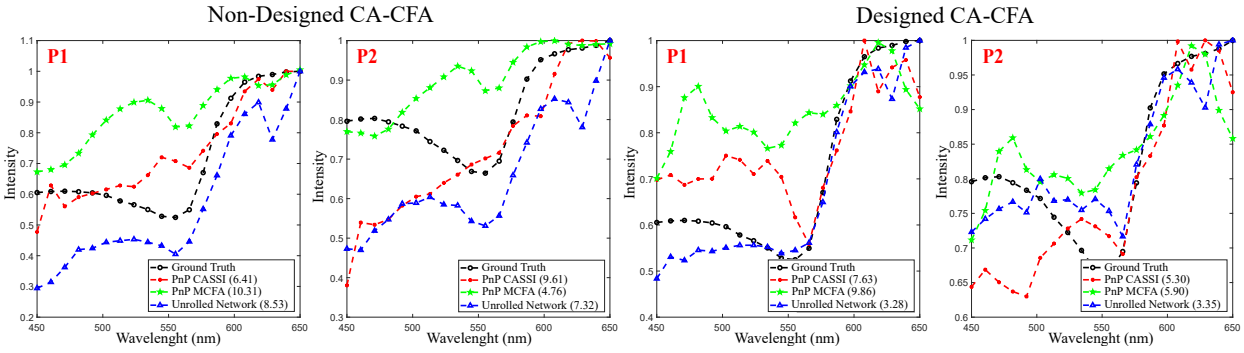

Figure 27. Reconstructed spectral signature for two random points in the image. The ground truth was sensed illuminated band-to-band with a commercial monochromator. Also, the SAM metric is shown in parenthesis.

Figure. 26 shows the compressive measurements acquired in with the CSIF optical architecture and a false RGB representation of the reconstructions images. The PnP algorithm was used to recover the spectral image from either the MCFA or the CASSI measurements. The unrolling fusion network showed a significant improvement compared with the reconstruction obtained with the PnP method and was more remarkable using the trained sensing matrices measurements. Fig. 27 shows the spectral signature of two random pixels. The SAM metric is shown in parenthesis measures the angle between the reconstructed spectra with the ground-truth spectra. The spectral reconstruction shows a highlighted improvement with the reconstruction using the unrolling fusion network with the trained CA and CFA.

7. Conclusions

The fusion of two compressive spectral imaging sensors using an E2E optimization for the joint design of the CA in both systems and a reconstruction network was presented. Particularly, it was proposed an optimization-inspired unrolling network for the recovery process. The CA of the sensing systems were learned as weights in a DNN constraining the learning of these parameters to binary values for implementation purpose. Simulation results showed that the proposed method outperforms a previous CSIF method, also, the E2E optimization of the CA improves significantly the reconstruction quality. Additionally, the proposed loss function boosts the network performance compared with the traditional mean square error loss function. A testbed implementation of the CSIF system was presented which validated the proposed method showing a better performance than reconstructing the compressive measurements separately.

The developing of this work resulted in two conference articles. One already accepted for The 28th IEEE International Conference on Image Processing (IEEE - ICIP) which will be held in Anchorage-Alaska in the period September 19-22, 2021. In this article it was presented the algorithmic methodology of the E2E optimization and the recovery network with only simulation results. And the other was accepted to the XXIII ”Simposio de Imagen, Procesamiento de Señales y Visión Artificial”(STSIVA 2021) where it was presented the testbed implementation of the proposed E2E optimization for CSIF.

Bibliographic References

Afonso, M. V., Bioucas-Dias, J. M., and Figueiredo, M. A. (2010). An augmented lagrangian approach to the constrained optimization formulation of imaging inverse problems. *IEEE Transactions on Image Processing*, 20(3):681–695.

Amro, I., Mateos, J., Vega, M., Molina, R., and Katsaggelos, A. K. (2011). A survey of classical methods and new trends in pansharpening of multispectral images. *EURASIP Journal on Advances in Signal Processing*, 2011(1):1–22.

Arad, B. and Ben-Shahar, O. (2016). Sparse recovery of hyperspectral signal from natural rgb images. In *European Conference on Computer Vision*, pages 19–34. Springer.

Arad, B., Timofte, R., Ben-Shahar, O., Lin, Y.-T., and Finlayson, G. D. (2020). Ntire 2020 challenge on spectral reconstruction from an rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 446–447.

Arce, G. R., Brady, D. J., Carin, L., Arguello, H., and Kittle, D. S. (2014). Compressive coded aperture spectral imaging: An introduction. *IEEE Signal Processing Magazine*, 31(1):105–115.

Arguello, H. and Arce, G. R. (2012). Restricted isometry property in coded aperture compressive spectral imaging. In *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pages 716–719.

Arguello, H. and Arce, G. R. (2014). Colored coded aperture design by concentration of measure in compressive spectral imaging. *IEEE Transactions on Image Processing*, 23(4):1896–1908.

Arguello, H. and Arce, G. R. (2014). Colored coded aperture design by concentration of measure in compressive spectral imaging. *IEEE Transactions on Image Processing*, 23(4):1896–1908.

Arguello, H., Rueda, H., Wu, Y., Prather, D. W., and Arce, G. R. (2013). Higher-order computational model for coded aperture spectral imaging. *Appl. Opt.*, 52(10):D12–D21.

Bacca, J., Galvis, L., and Arguello, H. (2020). Coupled deep learning coded aperture design for compressive image classification. *Optics express*, 28(6):8528–8540.

Bacca, J., Gelvez, T., and Arguello, H. (2021). Deep coded aperture design: An end-to-end approach for computational imaging tasks. *arXiv preprint arXiv:2105.03390*.

Baek, S.-H., Ikoma, H., Jeon, D. S., Li, Y., Heidrich, W., Wetzstein, G., and Kim, M. H. (2021). Single-shot hyperspectral-depth imaging with learned diffractive optics.

Bayer, B. E. (1976). Color imaging array. US Patent 3,971,065.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202.

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

Blumensath, T. and Davies, M. E. (2009). Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274.

Boyd, S., Parikh, N., and Chu, E. (2011). *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc.

Camps-Valls, G., Tuia, D., Bruzzone, L., and Benediktsson, J. A. (2014). Advances in hyperspectral image classification: Earth monitoring with statistical learning methods. *IEEE Signal Processing Magazine*, 31(1):45–54.

Candes, E. and Romberg, J. (2007). Sparsity and incoherence in compressive sampling. *Inverse problems*, 23(3):969.

Candes, E. J. and Wakin, M. B. (2008). An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30.

Chan, S. H., Wang, X., and Elgendy, O. A. (2016). Plug-and-play admm for image restoration: Fixed-point convergence and applications. *IEEE Transactions on Computational Imaging*, 3(1):84–98.

Chang, J. and Wetzstein, G. (2019). Deep optics for monocular depth estimation and 3d object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10193–10202.

Correa, C. V., Arguello, H., and Arce, G. R. (2016). Spatiotemporal blue noise coded aperture design for multi-shot compressive spectral imaging. *JOSA A*, 33(12):2312–2322.

Datcu, M. and Seidel, K. (2005). Human-centered concepts for exploration and understanding of

earth observation images. *IEEE Transactions on Geoscience and Remote Sensing*, 43(3):601–609.

Denman, S., Lamb, T., Fookes, C., Chandran, V., and Sridharan, S. (2010). Multi-spectral fusion for surveillance systems. *Computers & Electrical Engineering*, 36(4):643–663.

Figueiredo, M. A., Nowak, R. D., and Wright, S. J. (2007). Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of selected topics in signal processing*, 1(4):586–597.

Fischer, C. and Kakoulli, I. (2006). Multispectral and hyperspectral imaging technologies in conservation: current research and potential applications. *Studies in Conservation*, 51(sup1):3–16.

Foucart, S. and Rauhut, H. (2013). *An Invitation to Compressive Sensing*, pages 1–39. Springer New York, New York, NY.

Galvis, L., Arguello, H., and Arce, G. R. (2015). Coded aperture design in mismatched compressive spectral imaging. *Appl. Opt.*, 54(33):9875–9882.

Gastal, E. S. and Oliveira, M. M. (2011). Domain transform for edge-aware image and video processing. In *ACM SIGGRAPH 2011 papers*, pages 1–12.

Gat, N. (2000). Imaging spectroscopy using tunable filters: a review. In *Wavelet Applications VII*, volume 4056, pages 50–64. International Society for Optics and Photonics.

Gelvez, T. and Arguello, H. (2020). Nonlocal low-rank abundance prior for compressive spectral image fusion. *IEEE Transactions on Geoscience and Remote Sensing*.

Gelvez, T., Rueda, H., and Arguello, H. (2017). Joint sparse and low rank recovery algorithm for compressive hyperspectral imaging. *Appl. Opt.*, 56(24):6785–6795.

Gomez, R. B., Jazaeri, A., and Kafatos, M. (2001). Wavelet-based hyperspectral and multispectral image fusion. In Roper, W. E., editor, *Geo-Spatial Image and Data Exploitation II*, volume 4383, pages 36 – 42. International Society for Optics and Photonics, SPIE.

Govender, M., Chetty, K., and Bulcock, H. (2007). A review of hyperspectral remote sensing and its application in vegetation and water resource studies. *Water Sa*, 33(2).

Gregor, K. and LeCun, Y. (2010). Learning fast approximations of sparse coding. In *Proceedings of the 27th international conference on international conference on machine learning*, pages 399–406.

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., and Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, 187:27 – 48. Recent Developments on Deep Big Vision.

Gupta, R. and Hartley, R. I. (1997). Linear pushbroom cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(9):963–975.

Hagen, N. A. and Kudenov, M. W. (2013). Review of snapshot spectral imaging technologies. *Optical Engineering*, 52(9):1 – 23.

Hien Van Nguyen, Banerjee, A., and Chellappa, R. (2010). Tracking via object reflectance using a hyperspectral video camera. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 44–51.

Hinojosa, C., Bacca, J., and Arguello, H. (2018). Coded aperture design for compressive spectral subspace clustering. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1589–1600.

Horé, A. and Ziou, D. (2010). Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369.

Jácome, R., López, C., Garcia, H., and Arguello, H. (2021). Deep learning-based object classification for spectral images. In Orjuela-Cañón, A. D., Lopez, J., Arias-Londoño, J. D., and Figueroa-García, J. C., editors, *Applications of Computational Intelligence*, pages 147–159, Cham. Springer International Publishing.

Jagalingam, P. and Hegde, A. V. (2015). A review of quality metrics for fused image. *Aquatic Procedia*, 4:133–142.

Jerri, A. J. (1977). The shannon sampling theorem—its various extensions and applications: A tutorial review. *Proceedings of the IEEE*, 65(11):1565–1596.

Kimmel, R. (1999). Demosaicing: image reconstruction from color ccd samples. *IEEE Transactions on image processing*, 8(9):1221–1228.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lapray, P.-J., Wang, X., Thomas, J.-B., and Gouton, P. (2014). Multispectral filter arrays: Recent advances and practical implementation. *Sensors*, 14(11):21626–21659.

Li, C., Balla-Arabé, S., and Yang, F. (2016). Embedded multi-spectral image processing for real-time medical application. *Journal of Systems Architecture*, 64:26 – 36. Real-Time Signal Processing in Embedded Systems.

Li, R., Zheng, Y., Wen, D., and Song, Z. (2017). A deep learning approach to real-time recovery for compressive hyper spectral imaging. In *2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC)*, pages 1030–1034.

Mejia, Y. and Arguello, H. (2018). Binary codification design for compressive imaging by uniform sensing. *IEEE Transactions on Image Processing*, 27(12):5775–5786.

Miao, X., Yuan, X., Pu, Y., and Athitsos, V. (2019). l-net: Reconstruct hyperspectral images from a snapshot measurement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4059–4069.

Monga, V., Li, Y., and Eldar, Y. C. (2021). Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44.

Mousavi, A. and Baraniuk, R. G. (2017). Learning to invert: Signal recovery via deep convolutional networks. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2272–2276. IEEE.

Ongie, G., Jalal, A., Metzler, C. A., Baraniuk, R. G., Dimakis, A. G., and Willett, R. (2020). Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory*, 1(1):39–56.

Palsson, F., Sveinsson, J. R., and Ulfarsson, M. O. (2017). Multispectral and hyperspectral image fusion using a 3-d-convolutional neural network. *IEEE Geoscience and Remote Sensing Letters*, 14(5):639–643.

Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239.

Qiao, M., Meng, Z., Ma, J., and Yuan, X. (2020). Deep learning for video compressive sensing. *APL Photonics*, 5(3):030801.

Qin, J., Chao, K., Kim, M. S., Lu, R., and Burks, T. F. (2013). Hyperspectral and multispectral imaging for evaluating food safety and quality. *Journal of Food Engineering*, 118(2):157–171.

Ramirez, A., Arguello, H., Arce, G. R., and Sadler, B. M. (2014). Spectral image classification from optimal coded-aperture compressive measurements. *IEEE Transactions on Geoscience and Remote Sensing*, 52(6):3299–3309.

Renza, D., Martinez, E., and Arquero, A. (2013). A new approach to change detection in multispectral images by means of ergas index. *IEEE Geoscience and Remote Sensing Letters*, 10(1):76–80.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Rueda, H., Arguello, H., and Arce, G. R. (2016). Compressive spectral testbed imaging system based on thin-film color-patterned filter arrays. *Appl. Opt.*, 55(33):9584–9593.

Sitzmann, V., Diamond, S., Peng, Y., Dun, X., Boyd, S., Heidrich, W., Heide, F., and Wetzstein, G. (2018). End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Transactions on Graphics (TOG)*, 37(4):1–13.

Sogabe, Y., Sugimoto, S., Kurozumi, T., and Kimata, H. (2020). Admm-inspired reconstruction network for compressive spectral imaging. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2865–2869.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Tian, C., Xu, Y., Fei, L., and Yan, K. (2018). Deep learning for image denoising: a survey. In *International Conference on Genetic and Evolutionary Computing*, pages 563–572. Springer.

Vane, G., Green, R. O., Chrien, T. G., Enmark, H. T., Hansen, E. G., and Porter, W. M. (1993). The airborne visible/infrared imaging spectrometer (aviris). *Remote Sensing of Environment*, 44(2):127 – 143. Airbone Imaging Spectrometry.

Vargas, E., Arguello, H., and Tourneret, J.-Y. (2017). Spectral image fusion from compressive measurements using spectral unmixing. In *2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 1–5. IEEE.

Vargas, E., Espitia, O., Arguello, H., and Tourneret, J.-Y. (2018). Spectral image fusion from compressive measurements. *IEEE Transactions on Image Processing*, 28(5):2271–2282.

Wagadarikar, A., John, R., Willett, R., and Brady, D. (2008). Single disperser design for coded aperture snapshot spectral imaging. *Appl. Opt.*, 47(10):B44–B51.

Wang, L., Sun, C., Fu, Y., Kim, M. H., and Huang, H. (2019a). Hyperspectral image reconstruction using a deep spatial-spectral prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8032–8041.

Wang, L., Sun, C., Fu, Y., Kim, M. H., and Huang, H. (2019b). Hyperspectral image reconstruction using a deep spatial-spectral prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8032–8041.

Wang, L., Sun, C., Zhang, M., Fu, Y., and Huang, H. (2020a). Dnu: Deep non-local unrolling for computational spectral imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, L., Sun, C., Zhang, M., Fu, Y., and Huang, H. (2020b). Dnu: Deep non-local unrolling for computational spectral imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1661–1671.

Wang, L., Zhang, T., Fu, Y., and Huang, H. (2019). Hyperreconnet: Joint coded aperture optimization and image reconstruction for compressive hyperspectral imaging. *IEEE Transactions on Image Processing*, 28(5):2257–2270.

Wang, W., Zeng, W., Huang, Y., Ding, X., and Paisley, J. (2019). Deep blind hyperspectral image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4150–4159.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Wang, Z., Chen, B., Lu, R., Zhang, H., Liu, H., and Varshney, P. K. (2020c). Fusionnet: An unsupervised convolutional variational network for hyperspectral and multispectral image fusion. *IEEE Transactions on Image Processing*, 29:7565–7577.

Wang, Z., Simoncelli, E. P., and Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee.

Wei, Q., Bioucas-Dias, J., Dobigeon, N., and Tourneret, J.-Y. (2015). Hyperspectral and multispectral image fusion based on a sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 53(7):3658–3668.

Xie, Q., Zhou, M., Zhao, Q., Meng, D., Zuo, W., and Xu, Z. (2019). Multispectral and hyperspectral

image fusion by ms/hs fusion net. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1585–1594.

Xie, Q., Zhou, M., Zhao, Q., Xu, Z., and Meng, D. (2020). Mhf-net: an interpretable deep network for multispectral and hyperspectral image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xiong, Z., Shi, Z., Li, H., Wang, L., Liu, D., and Wu, F. (2017). Hscnn: Cnn-based hyperspectral image recovery from spectrally undersampled projections. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 518–525.

Yang, J., Zhao, Y., and Chan, J. (2018a). Hyperspectral and multispectral image fusion via deep two-branches convolutional neural network. *Remote Sensing*, 10:800.

Yang, J., Zhao, Y.-Q., and Chan, J. C.-W. (2018b). Hyperspectral and multispectral image fusion via deep two-branches convolutional neural network. *Remote Sensing*, 10(5):800.

Yokoya, N., Yairi, T., and Iwasaki, A. (2012). Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 50(2):528–537.

Yuan, X. (2016). Generalized alternating projection based total variation minimization for compressive sensing. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2539–2543.

Yuan, X., Liu, Y., Suo, J., and Dai, Q. (2020). Plug-and-play algorithms for large-scale snapshot compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1447–1457.

Yuhas, R. H., Goetz, A. F., and Boardman, J. W. (1992). Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In *Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop*, volume 1, pages 147–149.

Zhang, L., Zhang, Y., Yan, H., Gao, Y., and Wei, W. (2018). Salient object detection in hyperspectral imagery using multi-scale spectral-spatial gradient. *Neurocomputing*, 291:215 – 225.

Zhao, H., Gallo, O., Frosio, I., and Kautz, J. (2017). Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57.

Zhou, F., Hang, R., Liu, Q., and Yuan, X. (2019). Pyramid fully convolutional network for hyperspectral and multispectral image fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(5):1549–1558.

Zhu, R., Tsai, T.-H., and Brady, D. J. (2013). Coded aperture snapshot spectral imager based on liquid crystal spatial light modulator. In *Frontiers in Optics 2013*, page FW1D.4. Optical Society of America.

Zhuang, H., Deng, K., Fan, H., and Yu, M. (2016). Strategies combining spectral angle mapper and change vector analysis to unsupervised change detection in multispectral images. *IEEE Geoscience and Remote Sensing Letters*, 13(5):681–685.