Super resolution phase retrieval algorithm based on sparse priors

Jorge Luis Bacca Quintero

Doctoral thesis to qualify for the title of Doctor of Philosophy in Computer Science

Advisor:

Ph.D Henry Arguello Fuentes

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Ingenierías de Sistemas e Informatica

Bucaramanga

2021

Dedicatoria

I dedicate my dissertation to my family, friends, and parents.

Thank you for your unconditional support.

To Isabel to be my life partner

Agradecimientos

This work would not have been possible without the support, criticism, and help of my advisor Henry Arguello and the expert co-author and colleagues, among whom I wish to mention Samuel Pinilla, Claudia Correa, Laura Galvis, Carlos Hinojosa, Jesus Pineda, and Andres. G. Marrugo. Also, the numerous suggestions which came from the anonymous reviewers of the above publications. Thanks to the UIS and the HDSP group for taking me as a student and giving me all the means to become a Colombian Ph.D.

Table of Contents

List of Figures

List of Tables

Resumen

Título: Algoritmo de recuperación de fase de súper resolución basada en información de escasas [*]

Autor: Jorge Luis Bacca Quintero [**]

Palabras Clave: Super-resolución, Aperturas codificadas, Patrones difractivos codificados, Escases.

Descripción: La recuperación de la fase de alta resolución RFAR es un problema matemático inverso presente en imágenes óptica difractivas, el cual consiste en estimar una imagen de alta resolución a partir de medidas sin fase de baja resolución. Esta tesis estudia RFAR en un sistema óptico de patrones difractivos codificados, el cual introduce una apertura codificada (AC) para modular la fase, permitiendo adquirir multiples proyecciones desde el mismo objeto. Esta tesis doctoral considera dos escenarios de superresolución (i) computacional, donde las características del sensor determinan la resolución de la imagen recuperada, es decir, el tamaño de píxel del sensor es menor que el del AC, y (ii) físico, donde la resolución de la imagen está determinada por la resolución de la AC, asumiendo que el tamaño de píxel de AC es menor que la del sensor. Además, la estructura espacial de las AC puede diseñarse para mejorar la calidad de la estimación por lo tanto se desarrollan diferentes estrategias de diseño. Po otro lado, la literatura en algoritmos de recuperación han demostrado que las formulaciones no convexas superan los métodos convexos, requiriendo menos mediciones y complejidad computacional para recuperar la imagen. Sin embargo, la mayoría de los métodos no convexos se basan en una función de pérdida no suave y no incluyen información previa sobre la señal, como los escases. Por lo tanto, esta tesis estudia una función objetivo de mínimos cuadrados no convexos suavizada, donde se incluye algunos conocimientos previos sobre la señal, como escases, variación total y aprendizaje de los datos. Los resultados de la simulación muestran que los esquemas propuestos superan los métodos más avanzados

---

[*] Tesis de doctorado

[**] Facultad de Ingeniería Fisicomecánicas. Escuela de Ingeniería de Sistemas e Informática,
Director Henry Arguello Fuentes

en la reconstrucción de la imagen de alta resolución. Esta tesis también muestra que la calidad de la reconstrucción utilizando AC diseñada es superior a la de los conjuntos no diseñados.

Abstract

Title: Super resolution phase retrieval algorithm based on sparse priors. *

Author: Jorge Luis Bacca Quintero **

Keywords: Super resolution phase retrieval, Coded diffraction patterns, Coded aperture, Sparsity priors.

Description: Super-resolution phase retrieval (SRPR) is an inverse problem that appears in diffractive optical imaging and consists in estimating a high resolution image from low-resolution phaseless measurements. This thesis studies SRPR under a setup known as coded diffraction patterns, which introduces a coded aperture (CA) as a phase modulator encoding the diffraction patterns, allowing several projections from the same object. This doctoral thesis considers two super-resolution scenarios (i) computational, where the sensor characteristics mainly govern the attainable resolution of the recovered image, i.e., the pixel size of the sensor is smaller than that of the CA, and (ii) physical, the attainable resolution of the image is determined by the resolution of CA, assuming that the pixel size of CA is smaller than sensor pixel size. Additionally, the spatial structure of the CA can be designed to improve the quality of the estimation. Therefore, different strategies to design this spatial distribution are developed. From the recovery point of view, recent literature has shown that the non-convex formulations overcome the convex methods, requiring fewer measurements and less computational complexity to retrieve the phase image. However, most non-convex methods are based on non-smooth loss function, and they do not include prior information about the signal, such as sparsity. Therefore, this thesis studies a smoothed non-convex least-squares objective function, where some prior knowledge about the signal, such as sparsity, total-variation, and deep prior, is also included in the proposed formulation. Simulation results show that the proposed schemes overcome state-of-the-art methods in reconstructing the high-resolution image. This thesis also

---

* Super resolution phase retrieval algorithm based on sparse priors

** Facultad de Ingeniería Fisicomecánicas. Escuela de Ingeniería de Sistemas e Informática.
Advisor Henry Arguello Fuentes

shows that the reconstruction quality using designed CA is higher than that of the non-designed ensembles.

Introduction

In many science and engineering applications, it is required to estimate the phase of a complex signal from a set of intensity measurements. This problem is known as phase retrieval (PR), and it occurs in crystallography Pinilla et al. (2018b), Fresnel holography Poon and Liu (2014), lens-less imaging Shimano et al. (2018), astronomical imaging Fienup and Dainty (1987), and microscopy Dürig et al. (1986), among others. In particular, PR in optical imaging is needed since the electromagnetic diffracted field produced by an object when a laser beam illuminates it oscillates at rates of $\sim 10^{15}$ Hz, a rate which is impossible to be followed with current electronic measurement devices Shechtman et al. (2015). Indeed, traditional detectors (e.g., charge-coupled device (CCD) cameras, photosensitive films, and the human eye) measure the photon flux or irradiance, which are proportional to the magnitude squared of the field not the phase.

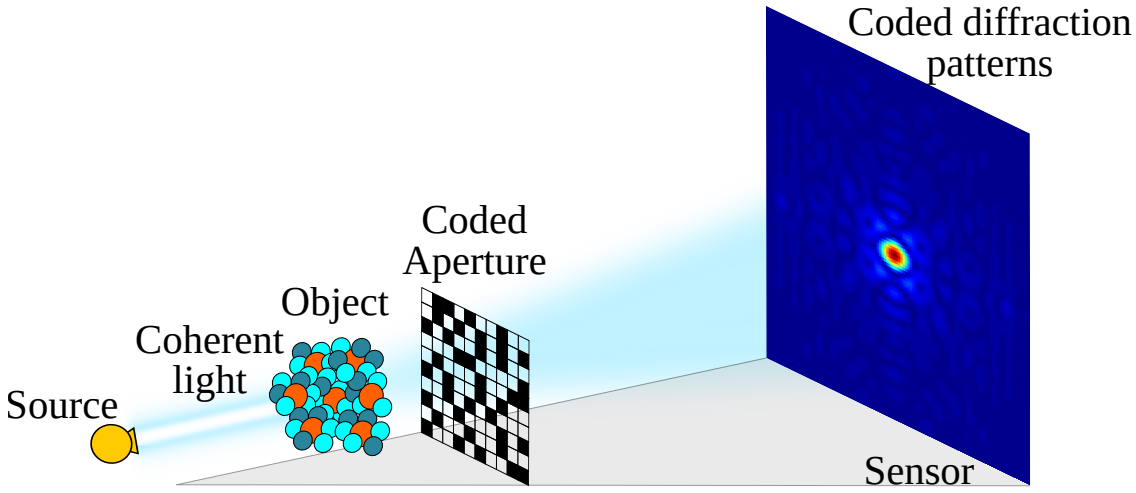
Alternatively, different approaches have been proposed to encode and then, with computation recovery algorithms, retrieve the phase, including holography Poon and Liu (2014), oversampling Fourier Miao et al. (2000), short time Fourier Griffin and Lim (1984) and coded diffraction patterns (CDP) Candes et al. (2015a), with the latter being the object of study of this work. CDP has attracted attention since it modifies traditional diffractive optical imaging by introducing an optical element known as a coded aperture which modulates the object diffracted field before the intensity of its diffraction pattern is sensed Pinilla et al. (2018b).

Figure 1 illustrates the optical setup to acquire CDP. Changing the spatial configuration of the coded aperture allows capturing multiple coded versions of the diffracted field of the scene.

Furthermore, this modulation provides uniqueness guarantees (up to a uni-modular constant) for a particular class of coded apertures Candes et al. (2015b); Shechtman et al. (2015), which has not been possible with traditional sensing systems. In general, modulations of this type can be attained in numerous ways: using a phase mask or an optical grating to modulate the illumination beam as mentioned in Loewen and Popov (2018), even by techniques from ptychography which scan different illumination angles on an extended specimen Rodenburg (2008); Thibault et al. (2009). Usually, the CA spatial distribution is chosen randomly; however, some recent works in other areas as tomography Mojica et al. (2017), compressive spectral imaging Arguello and Arce (2014) have improved the reconstruction quality by designing the spatial distribution. Therefore, this is a topic of interest for phase retrieval, and it is addressed in this thesis.

The main impediment to obtain a high-quality phase image from CDP measurements comes from the optics and the reconstruction algorithm. Notably, the optics impose physical constraints,

Figure 1. Schematic representation of a system that acquires coded diffraction patterns.



Note: A coded aperture is located after the object to modulate the field.

which limit the spatial resolution of the complex object; for instance, the low-pass filtering imposed by the propagation operator and by the pixelated discrete sensor Katkovnik et al. (2017). Although the spatial resolution of the sensor can be physically improved, the number of photons per unit area must increase to provide a reasonable signal-to-noise ratio (SNR), which results in longer exposure time Shechtman et al. (2015). Therefore, the need arises to design methodologies that allow obtaining a high-resolution phase image.

On the other hand, recently, phase retrieval algorithms with strong mathematical support have been proposed, which includes convex Candes et al. (2013, 2015a) and non-convex Chen and Candes (2015a) programming. Recent literature has shown that the non-convex formulations outperform the convex methods, requiring fewer measurements and less computational complexity to retrieve an image Pinilla et al. (2018a); Wang et al. (2018a,b); Chen and Candes (2015a); Candes et al. (2015d); Zhang et al. (2017). Despite the satisfactory performance of reconstruction algorithms to solve the phase retrieval problem, most non-convex methods are based on non-smooth loss function and they are not able to efficiently exploit signal properties such as sparsity, which has proven to be a powerful tool to reduce the number of measurements required to obtain a successful recovery phase image and also to increase the spatial resolution, resolving features smaller than one-fifth of the wavelength Szameit et al. (2012).

0.1. Scope of the Thesis

Therefore, this thesis focuses on three main aspects:

- Obtain a high-resolution image from low-resolution CDP. For that, this thesis proposed a new super-resolution scenario that this thesis denominated as *physical* since the spatial resolution

of a diffractive object can be determined by the resolution of the coded aperture instead of the sensor characteristic. This thesis establishes that an image can be recovered (up to a global uni-modular constant) with a high probability for this model.

- CA design to improve the quality of the reconstruction. As is shown in this thesis, the theoretical result states that the recovery probability from CDP directly depends on the CA, which can be increased by designing the CA. Therefore, a greedy strategy that designs the CA spatial distribution to maximize this probability is developed here. In addition, other design strategies are studied, such as end-to-end formulations that fit the network weights and the CA in a coupled manner.

- Development of new recovery algorithms. This thesis studies a new non-convex formulation based on the smoothing function, which overcomes the traditional non-smooth formulation. Additionally, sparse priors in the optimization algorithm and the initialization are explored to enhance the resolution and decrease the number of measurements necessary to retrieve the phase from CDP. Finally, recent deep priors are also studied to improve the reconstruction quality.

0.2. Publications and author's contribution

Most of the material presented in this thesis appears in the following publications by the author:

Journal Papers:.

1. Jorge Bacca, Yesid Fonseca,and Henry Arguello. Çompressive Spectral Image Reconstruc-

tion using Deep Prior and Low-Rank Tensor Representation". Accepted in Applied Optics (2021).

2. Jorge Bacca, Tatiana Gelvez and Henry Arguello. "Deep Coded Aperture Design: An End-to-EndApproach for Computational Imaging Tasks..Accepted in IEEE Transaction on Computational Imaging (2021).

3. Jorge Bacca, Laura Galvis, and Henry Arguello. " Coupled deep learning coded aperture design for compressive image classification". Optics Express (2020).

4. Jesus Pineda, Jorge Bacca, Jhacson Meza, Lenny Romero, Henry Arguello, and Andres G Marrugo. "SPUD: simultaneous phase unwrapping and denoising algorithm for phase imaging". Applied Optics (2020).

5. Jorge Bacca, Samuel Pinilla, and Henry Arguello. "Super-Resolution Phase Retrieval from Designed Coded Diffraction Patterns". IEEE Transactions on Image Processing (2019).

6. Jorge Bacca, Claudia Correa, and Henry Arguello, "Noniterative hyperspectral image reconstruction from compressive fused measurements," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 12, pp. 1231,1239, (2019).

7. Carlos Hinojosa, Jorge Bacca, and Henry Arguello, "Coded aperture design for compressive spectral subspace clustering," IEEE Journal of Selected Topics in Signal Processing, (2018).

8. Samuel Pinilla, Jorge Bacca, and Henry Arguello. "Phase retrieval algorithm via nonconvex minimization using a smoothing function." IEEE Transactions on Signal Processing 66.17

(2018): 4574-4584.

9. Jorge Bacca, Hector Vargas, H. M., Daniel Molina, and Henry Arguello. "Single pixel compressive spectral polarization imaging using a movable micro-polarizer array". Revista Facultad de Ingeniería Universidad de Antioquia, (88), 91-99 (2018).

Main Conference Papers:.

1. Jorge Bacca, Samuel Pinilla, Henry Arguello, (2019, September). Coded Aperture Design for Super-Resolution Phase Retrieval. In 2019 27th European Signal Processing Conference (EUSIPCO).
2. Samuel Pinilla, Jorge Bacca, Daniel Molina, Ariolfo Camacho, and Henry Arguello. (2018, June). Sparse Phase Retrieval Algorithm via Smoothing Function in Compressive Optical Imaging. In Mathematics in Imaging (pp. MW2D-2). Optical Society of America.
3. Jorge Bacca, Samuel Pinilla, Daniel Molina, Ariolfo Camacho, and Henry Arguello (2018, June). Super-Resolution Phase Retrieval Algorithm using a Smoothing Function. In Mathematics in Imaging (pp. MW2D-3). Optical Society of America.
4. Samuel Pinilla, Jorge Bacca, Jhon Angarita, and Henry Arguello (2018, April). Phase Retrieval via Smoothing Projected Gradient Method. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
5. Samuel Pinilla, Jorge Bacca, Jean-Yves Tourneret, and Henry Arugello (2018, June). A Smoothing Stochastic Phase Retrieval Algorithm for Solving Random Quadratic Systems. In 2018 IEEE Statistical Signal Processing Workshop (SSP).

Under Reviewer and other Publications.

1. Samuel Pinilla , Jorge Bacca, and Henry Arguello. "Sprsf: Sparse phase retrieval via smoothing function." arXiv preprint arXiv:1807.09703.
2. Samuel Pinilla , Jorge Bacca, Cesar Vargas, Juan Poveda, and Henry Arguello. Exact Crystalline Structure Recovery in X-ray Crystallography from Coded Diffraction Patterns. arXiv preprint arXiv:1907.03547.
3. Karen Sánchez, Jorge Bacca, Laura Arévalo, Henry Arguello,and S Castillo. Classification of Cocoa Beans Based on their Level of Fermentation using Spectral Information. TecnoLógicas, 24(50), e1654-e1654(2021).
4. Jorge Bacca, Henry Arguello. "Sparse Subspace Clustering for Hyperspectral Images using Incomplete Pixels". TecnoLogicas (2019)
5. Emmanuel Martínez, Santiago Castro, Jorge Bacca, J., and Henry. Transfer Learning for Spectral Image Reconstruction from RGB Images...Applications of Computational Intelligence: Third IEEE Colombian Conference, ColCACI 2020, Cali, Colombia, August 7-8, 2020, Revised Selected Papers 3. Springer International Publishing, (2021).

1. Objectives

General objective

- To design and implement an algorithm to reconstruct a high-resolution phase image from low-resolution quadratic measurements based on sparse representations.

Specific objectives

- To mathematically derive the super-resolution models in the phase retrieval problem to validate the assumption that a high-resolution phase can be recovered from low-resolution phaseless measurements.
- To establish a super-resolution discrete sensing model of the phase retrieval acquisition process from coded diffraction patterns.
- To develop a computational algorithm to simulate the super-resolution model from coded diffraction patterns.
- To derive theoretical guarantees to recover a high-resolution image from low-resolution measurements using approximate isometry properties.
- To formulate and design a numerical optimization problem which in- cludes the sparsity regularization term to reduce the number of measurements needed in the phase retrieval inverse problem.

- To evaluate the performance of retrieving the phase which the de- signed reconstruction algorithm compared with state-of-the-art methods.

2. Theoretical Background

Generally, the PR problem is formulated as recovering a signal $\mathbf{x} \in \mathbb{C}^n$ from a set of m quadratic equations of the form

$$y_k = |\langle \mathbf{a}_k, \mathbf{x} \rangle|^2, k = 1, \cdots, m, \tag{1}$$

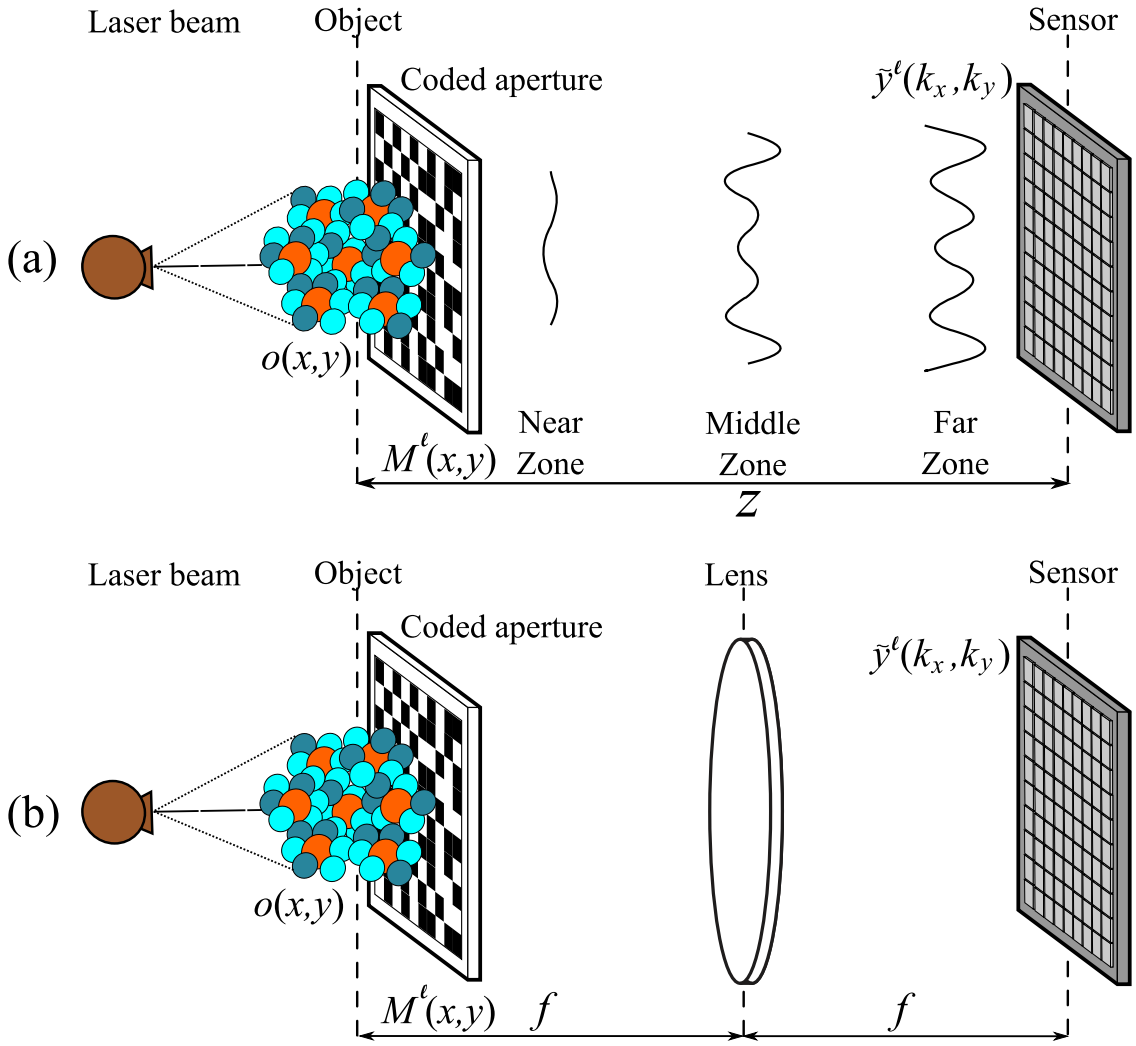
where $\mathbf{a}_k \in \mathbb{C}^n$ are the known sensing vectors and $\mathbf{y} = [y_1, \cdots y_m]^T$ is the phaseless measurement vector. For instance, the well-known Fourier phase retrieval, which corresponds to recover a signal from the modulus of its Fourier transform Candes et al. (2015a), occurs when $\mathbf{a}_k = [1, e^{-j2\pi k/m}, \cdots, e^{-j2\pi k(n-1)/m}]$, with $j = \sqrt{-1}$. PR is inherently an ill-posed problem, where due to the quadratic form, many signals may share the same magnitude measurements Shechtman et al. (2015). For mathematical analysis, $\mathbf{a}_k$ are often treated as random vectors, which under this assumption, the well-known trivial ambiguities are outperformed and uniqueness guarantees can be proven (up to a global constant), that are otherwise difficult to obtain Candes et al. (2015a); Gross et al. (2017); Candes et al. (2013). Specifically, any feasible solutions are given by $\hat{\mathbf{x}} = e^{j\theta}\mathbf{x}$, for some $\theta \in [0, 2\pi)$. In consequence, the euclidean distance between two complex vectors $\mathbf{w}_1$, $\mathbf{w}_2$ $\in \mathbb{C}^n$, on this phase retrieval problem is invariant to a global constant expressed as

$$dist(\mathbf{w}_1, \mathbf{w}_2) = \min_{\theta \in [0, 2\pi)} \|\mathbf{w}_1 e^{-j\theta} - \mathbf{w}_2\|_2. \tag{2}$$

One way to achieve the mathematical advantages provided by random sensing variables is by using a random coded aperture placed after the object to modulate its diffraction pattern and generate now a coded diffraction pattern that can be captured by a detector array.

2.1. Coded diffraction patterns (CDP)

Figure 2. Optical setups to obtained coded diffraction patterns.



Note: (a) Lens-less imaging and (b) $2f$-optical systems.

Several optical setups can be built to acquire CDP; for instance, two configurations are

illustrated in Fig. 2, where it can be noted that the introduced coded aperture allows acquiring of multiple projections of the object. In fact, these acquisition systems allow the acquisition of distinct measurements of the same scene by changing the spatial configuration of the coded aperture. Specifically, Fig. 2(a) illustrates a lens-less diffractive imaging system Goodman (2005). On the other hand, Fig. 2(b) represents a $2f$-optical setup, in which the lens adds phase to the diffraction patterns that are then recorded by the sensor. The standard formalization of the phase retrieval problem from coded diffraction patterns is of the form

$$y^{\ell}(k_x,k_y) = \left|\mathscr{A}\left(M^{\ell}(x,y)o(x,y)\right)\right|^2, \ell = 1,\cdots,L, \tag{3}$$

where $o(x,y)$ represents the object of interest, $M^{\ell}(x,y)$ models the ℓ-th configuration of the coded aperture, $\mathscr{A}$ is the complex-valued wavefront propagation operator from the target to the sensor plane which depending on the distance can be modeled as near, middle and far zone as described below, $y^{\ell}(k_x,k_y)$ are the phaseless measurements, L is the number of snapshots/projections, and (x,y), (k_x,k_y) the spatial and frequency coordinates, respectively.

According to the classical diffraction theory Poon and Liu (2014), the wavefront propagation from the object to a distance z, just before being measured, is modeled using three diffraction zones, known as near, middle and far fields, that are determined according to the distance between the object and the sensor Goodman (2008). Figure 2(a) shows a lensless diffractive imaging system that illustrates the three traditional diffraction zones. Mathematically, the coded diffraction patterns measured by the sensor at the different zones are as follows:

(a) Near zone:

$$y^{\ell}(x,y) = \mathscr{F}^{-1}\left\{T(k_x,k_y;z)\cdot\mathscr{F}\{M^{\ell}(x,y)o(x,y)\}\right\}, \tag{4}$$

where $T(k_x,k_y;z)$ denotes the spatial frequency transfer function defined as

$$T(k_x,k_y;z) = \ e^{\frac{j2\pi z}{\lambda}\sqrt{1-\lambda^2(k_x^2+k_y^2)}}, \tag{5}$$

assuming $k_x^2+k_y^2 \leq \frac{1}{\lambda^2}$ Poon and Liu (2014). This zone is considered in applications such as optical microscopy Dürig et al. (1986), near-field raman imaging Jahncke et al. (1995), and near-field spectroscopy Hess et al. (1994).

(b) Middle zone :

$$y^{\ell}(k_x,k_y) \propto e^{\left(\frac{j\pi(k_x^2+k_y^2)}{\lambda z}\right)}\mathscr{F}\left\{M^{\ell}(x,y)o(x,y)e^{\left(\frac{j\pi(x^2+y^2)}{\lambda z}\right)}\right\}. \tag{6}$$

Applications such as Fresnel holography Poon and Liu (2014) and lens-less imaging Shimano et al. (2018) are based on the middle diffraction zone to develop new acquisition imaging devices Shimano et al. (2018).

(c) Far zone:

$$y^{\ell}(k_x,k_y) \propto \mathscr{F}\{M^{\ell}(x,y)o(x,y)\}. \tag{7}$$

The far diffraction zone has allowed the development of applications such as crystallography, astronomical imaging and microscopy Goodman (2008). Indeed, the system in Fig.2(b) is modeled with the far field.

For all the cases $y^{\ell}(k_x,k_y) = \mathscr{A}(M^{\ell}(x,y)o(x,y))$, λ is the wavelength of the incident light and $\mathscr{F}(\cdot)$, and $\mathscr{F}^{-1}(\cdot)$ are the Fourier and the inverse Fourier transform, respectively Goodman (2008).

2.1.1. ***Forward matrix model.*** Since the sensor is a finite two-dimensional pixel array, as illustrated in Fig. 2, the model should be discretized. Conventionally, the object, sensor, and coded aperture are assumed to have the same pixel size (Δ). So, it is convenient to introduce the following matrices $\mathbf{X},\mathbf{Q},\mathbf{M}_{\ell} \in \mathbb{C}^{N\times N}$ where N^2 stands for the number of pixels, which is given as follows:

$$\begin{aligned}
\mathbf{X}[x,y] &= o_{x,y},\\
\mathbf{M}_{\ell}[x,y] &= M^{\ell}_{x,y},\\
\mathbf{Q}[x,y] &= e^{\frac{j\pi z\lambda}{(N\Delta)^2}(x^2+y^2)},\\
\mathbf{T}[x,y] &= e^{\frac{-j2\pi z}{\lambda}\sqrt{1-\frac{\lambda^2(s^2+r^2)}{M^2\Delta_s}}}
\end{aligned} \tag{8}$$

where $\mathbf{X}[x,y],\mathbf{M}_{\ell}[x,y]$ represent the 2D discretization of the object and the coded aperture, respectively, $\mathbf{T}[x,y]$ and $\mathbf{Q}[x,y]$ are orthogonal diagonal matrices that model the discrete spatial frequency transfer function for the near and middle zone, respectively Goodman (2008); Poon and Liu (2014). Also, taking $\mathbf{x}$ as a column-wise vectorization of the matrix $\mathbf{X}$, and defining $\mathbf{F} \in \mathbb{C}^{N\times N}$ as the dis-

crete Fourier transform matrix with entries

$$\mathbf{f}_i^H = \frac{1}{\sqrt{N}}[\omega^{-0(i-1)}, \omega^{-1(i-1)}, \cdots, \omega^{-(n-1)(i-1)}], \tag{9}$$

such that $i = 1, \cdots, N$ and $\omega = e^{\frac{2\pi j}{N}}$ is the N-th root of unity; the acquired CDP at the three diffraction zones are given by

$$\begin{aligned} \mathbf{y}_\ell &= |\mathbf{FTF}^H\mathbf{M}_\ell\mathbf{x}|^2 + \eta_\ell && (\textit{Near zone}) , \\ \mathbf{y}_\ell &= |\mathbf{F}^H\mathbf{QM}_\ell\mathbf{x}|^2 + \eta_\ell && (\textit{Middle zone}) , \\ \mathbf{y}_\ell &= |\mathbf{FM}_\ell\mathbf{x}|^2 + \eta_\ell && (\textit{Far zone}) , \end{aligned} \tag{10}$$

where η_ℓ is the observation additive noise, for ℓ-th measurement.

2.2. Traditional Phase Retrieval Recovery Methods

This chapter section introduces some state-of-the-art algorithms developed to solve the phase retrieval problem. Traditional methods can be traced back to the 1070s, where Error-Reduction methods Fienup (1982) were proposed. These empirical methods iteratively apply projections between the signal and the phaseless measurements, where some prior knowledge about its structure is incorporated. However, these algorithms do not have theoretical guarantees of convergence Candes et al. (2015e); Fienup (1982). In recent years, modern convex and non-convex optimization approaches with theoretical guarantees of convergence and recovery have been proposed. All of them solve (1), which is the generalized phase retrieval problem. This thesis is focused on the

non-convex formulation where some recovery algorithms are developed, of which it is worth highlighting, the *smoothing phase retrieval*, where some variants based on sparsity are also proposed, and the *E2E approach* based on deep learning decoders which is present in Section 5.

2.2.1. *Convex Approaches*. Using the lifting trick, the complex signal can be expressed as a rank-one matrix and retrieved from quadratic measurements by solving semidefinite programming (SDP). Some algorithms under this approach are described below.

2.2.1.1. PhaseLift Candes et al. (2013). The phase retrieval problem can be seen as a problem to recover a low-rank matrix $\mathbf{X} = \mathbf{x}\mathbf{x}^H$. This problem was first addressed by the PhaseLift method in Candes et al. (2013), where it is shown that it is possible to recover exactly one signal from phaseless measurements contaminated with additive noise, solving the following optimization problem

$$\begin{aligned} &\arg\min_{\mathbf{X}\in\mathbb{S}^{n\times n}} \ Tr(\mathbf{X}) \\ &\text{subject to} \quad \begin{aligned} &\mathbf{X} \succeq 0 \\ &\|Tr(\mathbf{a}_k\mathbf{a}_k^H\mathbf{X}) - y_k\|_2 = \|\mathscr{A}(\mathbf{X}) - \mathbf{y}\|_2 \leq \varepsilon,\ k = 1,\cdots,m. \end{aligned} \end{aligned} \tag{11}$$

where $\mathscr{A}(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$ is a linear operator, $\mathbf{X} = \mathbf{x}\mathbf{x}^H$ and $\|\cdot\|_2$ is the euclidean norm. This problem is relaxed by following a minimization problem of the form

$$\arg\min_{\mathbf{X}\in\mathbb{S}^{n\times n}} \ \frac{1}{2}\|\mathscr{A}(\mathbf{X}) - \mathbf{y}\|_2 + \tau\|\mathbf{X}\|_*, \tag{12}$$

known as *Regularized Nuclear Norm*, in the standard theory of optimization Ekeland and Temam (1999).

2.2.1.2. Normally Distributed Candes et al. (2015c). Assuming that the measurements are a sequence of independent samples of the Gaussian distribution with mean u_k and variance σ_k^2, the optimization problem minimizing the log-likelihood function for independent Gaussian samples is given by

$$\begin{array}{ll} \arg\min_{\mathbf{X}\in\mathbb{S}^{n\times n}} & \sum_{k=1}^{m} \frac{1}{2\sigma_k^2}(y_k - u_k)^2 + \tau Tr(\mathbf{X}) \\ \text{subject to} & \mathbf{u} = \mathscr{A}(\mathbf{X}) \\ & \mathbf{X} \succeq 0 \end{array} . \tag{13}$$

Further if Σ is a diagonal matrix with diagonal elements σ_k^2, the optimization problem in (13) can be rewritten as

$$\begin{array}{ll} \arg\min_{\mathbf{X}\in\mathbb{S}^{n\times n}} & \frac{1}{2}(\mathbf{y} - \mathscr{A}(\mathbf{X}))^H \Sigma^{-1}(\mathbf{y} - \mathscr{A}(\mathbf{X})) + \tau Tr(\mathbf{X}) \\ \text{subject to} & \mathbf{X} \succeq 0 \end{array} , \tag{14}$$

where τ is a regularization parameter.

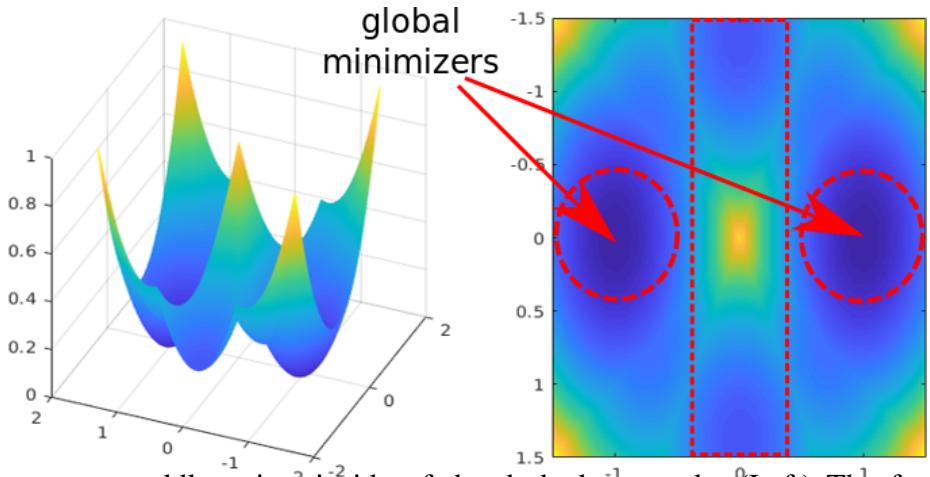
However, the lifting trick methods are not widely used because they are impractically expensive other than for small-scale problems. Therefore, some non-convex approaches are developed.

2.2.2. *Non-Convex Approaches*. Conventionally, the least-squares criterion has been adopted to recover $\mathbf{x}$, by minimizing the intensity-based empirical loss

$$\min_{\mathbf{x}\in\mathbb{C}^n} h(\mathbf{x}) = \sum_{k=1}^{m} \left(|\mathbf{a}_k\mathbf{x}|^2 - y_k\right)^2. \tag{15}$$

Due to the absolute value, (15) is highly non-convex. Fig. 3 plots the landscape of (15) for $\mathbf{x} = [0,1]^T$, and it can be seen that $\pm\mathbf{x}$ are global minimizers located inside the dashed circles. In other words, (15) is convex near to the global minimizers. Additionally, the dashed rectangle contains two saddle points that interfere with the convergence of any phase retrieval iterative process. Then, any algorithm that solves (15) can recover $\mathbf{x}$ if it can reach these convexity zones. In the literature, these regions are reached by performing a carefully statistical initialization of $\mathbf{x}$, which is then refined based on gradient-based methods using the Wirtinger derivativeCandes et al. (2015d). This

Figure 3. Landscape of (15) *for* $\mathbf{x} = [0,1]^T$ *and* $L = 4$.



Note: There are two saddle points inside of the dashed rectangle. (Left) The function graph; (Right) The same function visualized as a color image. The coded apertures take values in $d[t] \in \{1,-1,j,-j\}$. The dashed circles are the convexity regions of (15).

thesis is focused on the refinement step.

2.2.3. *Initialization*.

2.2.3.1. Weighted Maximal Correlation initialization. The Weighted Maximal Correlation initialization proposed in Wang et al. (2017b) is based on exploiting the orthogonality characteristics of the high-dimensional signal, which is assumed to be orthogonal with the random sampling vector. Specifically, the initialization consists in the weighted maximal correlation initialization proposed in Wang et al. (2017b), is based on exploiting the orthogonality characteristics of the high-dimensional signal, which is assumed to be orthogonal with the random sampling vector. Precisely, The initialization consists in calculating the vector $\mathbf{x}_0$, which is the leading eigenvector $\tilde{\mathbf{z}}_0$ of the matrix

$$\mathbf{Y}_0 := \frac{1}{|I_0|}\sum_{k\in I_0}\sqrt{q_k}\frac{\mathbf{a}_k\mathbf{a}_k^H}{\|\mathbf{a}_k\|_2^2}, \tag{16}$$

scaled by the quantity $\lambda_0 := \sqrt{\frac{\sum_{k=1}^m q_k^2}{m}}$, *i.e*, $\mathbf{x}_0 = \lambda_0\tilde{\mathbf{z}}_0$. The set I_0 is the collection of indices corresponding to the largest values of $\{|\langle\mathbf{a}_k,\mathbf{x}\rangle|/\|\mathbf{a}_k\|_2\}$. The notation $|I_0|$ is the cardinality of the set I_0 which are usually chosen as $\lfloor\frac{3m}{13}\rfloor$, where $\lfloor w\rfloor$ denotes the largest integer number less than w. Moreover, in Wang et al. (2016a) it was established that the distance between the initial guess $\mathbf{x}_0$ and the true signal $\mathbf{x}$ is given by

$$d_r(\mathbf{x}_0,\mathbf{x}) \leq \frac{1}{10}\|\mathbf{x}\|_2, \tag{17}$$

with probability exceeding $1-c_3e^{-c_4m}$, providing that $m \geq c_1|I_0| \geq c_2n$ for some constants $c_1,c_2,c_3,c_4 > 0$ and sufficiently large n.

Algorithm 1 Filtered Spectral Initialization (FSI)

1: **Input:** Acquired data $\{\mathbf{a}_k; y_k\}$, maximum number of iterations T, and a low pass filter $\mathscr{G}$.
2: $\tilde{\mathbf{x}}^{(0)} \leftarrow$ Chosen randomly.
3: **Set** $\overline{\mathscr{I}_0}$ as the set of indices corresponding to the $|\overline{\mathscr{I}_0}|$ largest values of $\{y_k/\|\mathbf{a}_k\|_2\}$, and build matrix $\mathbf{Y}_0$.
4: **for** $r = 0 : R-1$ **do**
5: $\quad \grave{\mathbf{x}}^{(r+1)} = \mathscr{G}\left(\mathbf{Y}_0 \tilde{\mathbf{x}}^{(r)}\right)$.
6: $\quad \tilde{\mathbf{x}}^{(r+1)} = \frac{\grave{\mathbf{x}}^{(r+1)}}{\left\|\grave{\mathbf{x}}^{(r+1)}\right\|_2}$.
7: **end for**
8: Compute $\mathbf{x}^{(0)} = \sqrt{\frac{\sum_{k=1}^{m} y_k}{m}}\tilde{\mathbf{x}}^{(R)}$.
9: **Return:** $\mathbf{x}^{(0)}$

2.2.3.2. Filtered Spectral Initialization. Considering that the scene initial estimation plays a vital role in solving the phase retrieval problem, this section presents one of the available initialization, the called filtered spectral initialization (FSI). This procedure takes advantage of the extension to CDP of the orthogonality-promoting initialization in Jerez et al. (2020). Roughly speaking, FSI calculates an estimate of the scene, $\mathbf{x}^{(0)}$, computing a low-pass version of the leading eigenvector $\tilde{\mathbf{x}}^{(0)}$ of the matrix

$$\mathbf{Y}_0 := \frac{1}{|\mathscr{I}_0|} \sum_{(\ell,k)\in\mathscr{I}_0} \frac{\mathbf{a}_k \mathbf{a}_k^H}{\|\mathbf{a}_k\|_2^2}, \tag{18}$$

scaled by the quantity $\lambda_0 := \sqrt{\frac{\sum_{k=1}^{m} y_k}{m}}$, i.e, $\mathbf{x}^{(0)} = \lambda_0 \tilde{\mathbf{x}}^{(0)}$, where the set $\mathscr{I}_0$ contains the values of s associated with the $\lfloor \frac{m}{2} \rfloor$ largest values of $\{y_k/\|\mathbf{a}_k\|_2\}$. The notation $|\mathscr{I}_0|$ represents the cardinality of the set $\mathscr{I}_0$, and $\overline{(\cdot)}$ is the complex conjugate operation. The low-pass constraint of FSI comes from the fact that an image is mostly composed of low frequencies Jerez et al. (2020). Since FSI requires the estimation of the leading eigenvector of $\mathbf{Y}_0$, the simplest way to achieve it is

following a power iteration strategy Wang et al. (2018a). In plain words, this strategy consists in recursively performing a matrix-vector multiplication between $\mathbf{Y}_0$ and the current estimation of the scene, as summarized in Algorithm 1. Precisely, in line 5, a low-pass filter $\mathscr{G}$ is applied to the current estimation of $\mathbf{x}$, $\mathbf{Y}_0\tilde{\mathbf{x}}^{(t)}$, to preserve the low frequency information. Here, it is worth mentioning that this filtering process is the crucial step of the initialization and the main difference with the orthogonality-promoting initialization (OPI), as illustrated in Fig. 2. In fact, Fig. 2 shows that performing the filtering step provides a closer estimation of the scene using fewer phaseless measurements. For this particular experiment, $\mathscr{G}$ was fixed as the Gaussian filter. Finally, algorithm 1 returns the scaled vector $\mathbf{x}^{(0)}$ as the estimation of the scene, in line 9. Given the initialization returned by Algorithm 1, it has been recently used to develop a rapid target detection method from CDP achieving high detection rates using the minimal number of measurements Jerez et al. (2020).

2.2.4. *Refinement Step*.

2.2.4.1. Truncated Wirtinger Flow (TWF) Chen and Candes (2015a). Suppose that the measurements are a sequence of independent samples of the Poisson distributions, *i.e* $y_k \sim Poisson(|\langle \mathbf{a}_k, \mathbf{x}\rangle|^2)$. Calculating the log-Likelihood for independent samples, has the form $\sum_{k=1}^{m} y_k log(u_k) - u_k$, where $\mathbf{u} = \mathscr{A}(\mathbf{X})$, then the optimization problem to recover the phase can be written as

$$\begin{array}{ll} \arg\min_{\mathbf{X}\in\mathbb{S}^{n\times n}} & \sum_{k=1}^{m} y_k log(u_k) - u_k + \tau Tr(\mathbf{X}) \\ \text{subject to} & \mathbf{u} = \mathscr{A}(\mathbf{X}) \\ & \mathbf{X} \succeq 0 \end{array} , \tag{19}$$

where τ is a regularization parameter.

2.2.4.2. Truncated Amplitude Flow (TAF) Wang et al. (2018a). Adopting the least-squares criterion, the task of recovering a solution from the phaseless measurements reduces to that of minimizing the amplitude-based loss function

$$\min_{\mathbf{x}\in\mathbb{C}^n} f(\mathbf{x}) = \frac{1}{m}\sum_{k=1}^{m}\left(|\langle \mathbf{a}_k, \mathbf{x}\rangle| - q_k\right)^2, \tag{20}$$

where and $q_k = \sqrt{y_k}$. Notice that the optimization problem in (20) is non-smooth and non-convex. To solve these issues a smoothing phase retrieval algorithm has been developed; it will be introduced in the following.

2.2.4.3. Sparse priors. Some prior knowledge of the object can be incorporated into the phase retrieval problem to help regularize, which is the case of a real-space object that is sparse in some known representation Baraniuk (2007). Mathematically, this means that an object is represented as a sparse linear combination on a basis as

$$\mathbf{x} = \Psi\theta, \tag{21}$$

where $\Psi \in \mathbb{C}^{n\times n}$ stands for the sparsity basis and θ is a sparse vector, i.e., $||\theta||_0 = k << n$, where vector contains a small number of non-zero coefficients. This constraint can be practically incorporated into convex and non-convex optimization problems.

2.2.4.4. SDP-Based Methods with sparsity prior. Since the result outer product $\mathbf{X} = \mathbf{x}\mathbf{x}^H$ of a sparse signal is a sparse matrix as well Ohlsson et al. (2011), SDP methods can incorporate this prior information by minimizing the ℓ_1 norm of the matrix $\mathbf{X}$. This formulation yields

$$\begin{aligned} \arg\min \quad & Tr(\mathbf{X}) + \tau \|\mathbf{X}\|_1 \\ & \left| Tr(\mathbf{a}_k \mathbf{a}_k^H \mathbf{X}) - \mathbf{y} \right| \leq \varepsilon, \; k = 1, \cdots, m \;, \\ \text{subject to} \quad & \mathbf{X} \succeq 0, \end{aligned} \tag{22}$$

where the solution of (22) is shown in Ohlsson et al. (2011) to be unique in the noiseless cases. Further, in Li and Voroninski (2013) it was shown that for independent zero-mean sensing vectors, on the order of $\mathscr{O}(k^2 \log(n))$ measurements are needed to recover a k-sparse vector. However, as explained in the previous section, the SDP methods are expensive to real problems where the dimensions of the data increase.

2.2.4.5. Wirtinger-Based Methods with sparsity prior. Recently, Wirtinger-based methods have been extended to PR with sparse inputs. Sparse WF (SWF) Yuan et al. (2019), and the sparse truncated amplitude flow (SPARTA) Wang et al. (2017c),use different initialization strategies to guarantee exact recovery of the true signal. In particular, SPARTA introduces the sparse orthogonality promoting initialization in Wang et al. (2017c), and SWF proposes a variant of the spectral initialization developed in Yuan et al. (2019). The initialization for the SWF algorithm returns a more accurate estimate of the true signal than the SPARTA initialization. On the other hand, to exploit the sparsity prior, in the refinement step, a k-sparse hard threshold operator $\mathscr{H}_k(\cdot)$is

iteratively applied in the gradient descent direction as

$$\mathbf{x}^{(t+1)} = \mathscr{H}_k\left(\mathbf{x}^{(t)} - \tau\partial f\left(\mathbf{x}^{(t)}\right)\right),\ t = 1,2\cdots \tag{23}$$

where $\mathscr{H}_k(\mathbf{u})$sets all the entries in the vector $\mathbf{u} \in \mathbb{C}^n$ to zero, except for its k largest absolute values.This procedure can decrease the freedom of dimensions, constraining the searching domain and converges linearly for any k-sparse n-long signal $(k \ll n)$ with sampling complexity $\mathscr{O}(k^2\log(n))$ Wang et al. (2017c); Yuan et al. (2019).

3. Super resolution phase retrieval problem

Before showing the super-resolution scenarios, it is worth describing the mathematical model of the coded aperture since it is an important optical element that allows obtaining the proposed super-resolution scenario that this thesis denominated as *physical.* This optical element is a two-dimensional array of square hardware pixels, as illustrated in Fig 4. Defining Δ_m as the pixel size of the coded aperture, its transmittance function can be expressed as

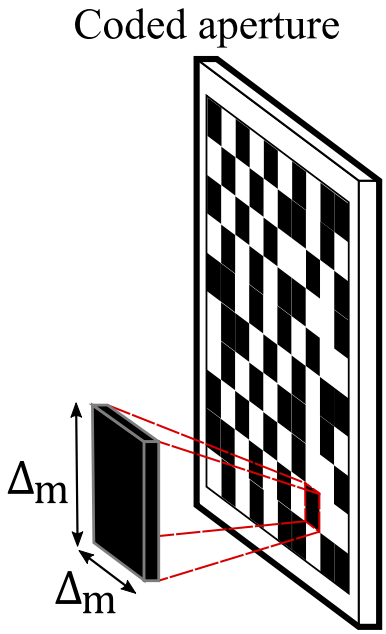
$$M(x,y) = \sum_{k',l'} M_{k',l'} \text{rect}\left(\frac{x}{\Delta_m} - k', \frac{y}{\Delta_m} - l'\right), \tag{24}$$

where $M^{\ell}_{k',l'}$ represents the value at the pixel indexed by k',l'. Notice that the coded aperture is described as a continuous model, but it has the same spatial value in an area of$\Delta_m \times \Delta_m$ given by the pixels of the coded aperture. The spatial resolution in phase retrieval is limited by two principal factors: low-pass filtering by the propagation operators and by the pixel size in the pixelated discrete sensor and the coded aperture Katkovnik et al. (2017), being the last one the case of study of this thesis.

Discretization of the encoded object. In CDP, the object and the CA are placed in the same plane, i.e., the wavefront just before the coded aperture, which is called the *encoded object* is described as

$$\tilde{o}^{\ell}(x,y) = M^{\ell}(x,y)o(x,y) \tag{25}$$

Figure 4. Visual representation of the coded aperture



Note: The pixel size is denoted as $(\Delta_m \times \Delta_m)$

where ℓ stands for the number of projections of the same object with the different spatial distribution of the CA. It is worth noting that the encoded object is continuous (although it passes through the CA that has discrete pixels), and the sensor is responsible for the discretization of the wavefront.

Therefore, assuming Δ_s as the pixel size of the sensor, the sampling period of the encoded object must satisfy the following equality Goodman (2005); Poon and Liu (2014)

$$\Delta_{\tilde{o}} = \frac{\lambda z}{S\Delta_s}, \tag{26}$$

where $S \times S$ is the square-support of the sensor, λ is the wavelength, and z is the propagation

distance. Moreover, as the object and the CA are in the same plane, considering (26) and (24), the discrete value of the encoded object $\tilde{o}^{\ell}(x,y)$ is modeled as

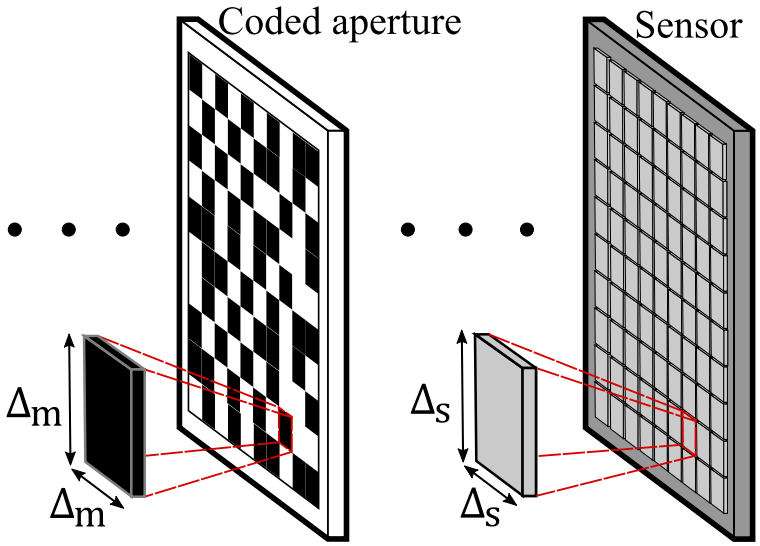
$$\begin{aligned}\tilde{o}^{\ell}_{k,l} &= \iint \text{rect}\left(\frac{x}{\Delta_{\tilde{o}}}-k,\frac{y}{\Delta_{\tilde{o}}}-l\right)\tilde{o}^{\ell}(x,y)dxdy \\ &= \iint \text{rect}\left(\frac{x}{\Delta_{\tilde{o}}}-k,\frac{y}{\Delta_{\tilde{o}}}-l\right)\times \\ &\quad \underbrace{\sum_{t,q} M^{\ell}_{t,q}\text{rect}\left(\frac{x}{\Delta_m}-t,\frac{y}{\Delta_m}-q\right)}_{M^{\ell}(x,y)} o(x,y)dxdy,\end{aligned} \tag{27}$$

for $k,l = 1,\cdots,S$. Considering this setup, and assuming the pixel size of the coded aperture ($\Delta_m \times \Delta_m$) and the sensor ($\Delta_s \times \Delta_s$) as is shown in Fig. 5, some scenarios of the discretization of the continuous object[1] $o(x,y)$ can be considered:

- $\Delta_o = \Delta_{\tilde{o}} = \Delta_m = \Delta_s$, which corresponds to the traditional phase retrieval problem.
- $\Delta_o = \Delta_s = \Delta_{\tilde{o}}$ and $\Delta_m > \Delta_s$, can be treated as the traditional phase retrieval problem with super-pixel in the coded aperture.
- $\Delta_o < \Delta_s = \Delta_{\tilde{o}} = \Delta_m$, this setup, is considered as a *computational super resolution scenario* Katkovnik et al. (2017), since according with the pixel size Δ_s it is not physically possible to reconstruct an image with higher spatial resolution. However, only a pure computational

[1] It is important to highlight that this discretization is computational and is necessary to perform employed phase algorithms. The computational pixel value can be as small as desired; however, different small features are achieved due to the optical elements.

Figure 5. Coded Aperture and Sensor



Note: Pixel size of the coded aperture ($\Delta_m \times \Delta_m$) and the sensor ($\Delta_s \times \Delta_s$)

model with a computation sampling period of Δ_o can be used in the recovery algorithm, and then decimated into the sensor.

- $\Delta_o = \Delta_m < \Delta_{\tilde{o}} < \Delta_s$, this setup considers a pixel size of the coded aperture Δ_m smaller than $\Delta_{\tilde{o}}$ which is the pixel size of the target image that the sensor allows. Therefore, the spatial resolution of the image depends on Δ_m instead of the pixel size of the sensor Δ_s, leading to a *physical super-resolution scenario*; for this reason Δ_m should be smaller than $\Delta_{\tilde{o}}$. Additionally, this setup is of high interest because it corresponds to a hardware super-resolution scenario.

3.1. Physical super resolution phase retrieval

This thesis focuses in the case when $\Delta_m < \Delta_{\tilde{o}}$, which is the scenario showed in Fig 6. In order to simplify (27), assume that $\Delta_{\tilde{o}} = r_m \Delta_m$, where $r_m \geq 1$ is an integer up-sampling factor,

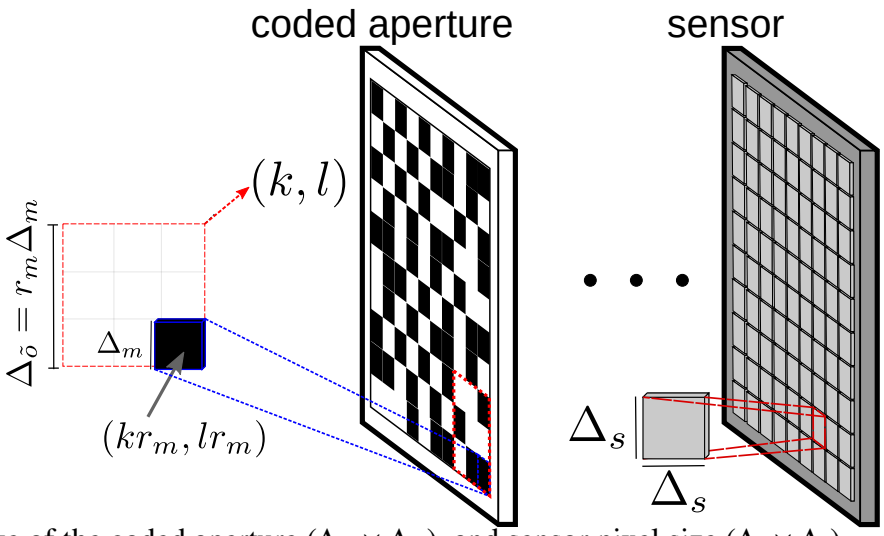
which means that within the squared section $\Delta_{\tilde{o}} \times \Delta_{\tilde{o}}$ there are $r_m \times r_m$ pixels of the coded aperture of size $\Delta_m \times \Delta_m$ as illustrated in Fig 5. This fact implies that the effect of rect $\left(\frac{x}{\Delta_{\tilde{o}}} - k, \frac{y}{\Delta_{\tilde{o}}} - l\right)$ in (27) can be equivalently expressed by limiting the indexing variables t, q as

$$1 + (k-1)r_m \leq t \leq kr_m$$
$$1 + (l-1)r_m \leq q \leq lr_m. \tag{28}$$

In consequence, (27) can be rewritten as

$$\tilde{o}^{\ell}_{k,l} = \sum_{t=1+(k-1)r_m}^{kr_m} \sum_{q=1+(l-1)r_m}^{lr_m} M^{\ell}_{t,q} o_{t,q} \tag{29}$$

Figure 6. Super-resolution scenario.



Note: Pixel size of the coded aperture ($\Delta_m \times \Delta_m$), and sensor pixel size ($\Delta_s \times \Delta_s$).

where $o_{t,q} = \iint \text{rect}\left(\frac{x}{\Delta_m} - t, \frac{y}{\Delta_m} - q\right) o(x,y)dxdy$ models the discrete values of the target image, for $t,q = 1,\cdots,r_m S$.

Thus, from (29), and the relationship $z \geq \frac{2S\Delta_o^2}{\lambda}$, presented in Poon and Liu (2014); Goodman (2005) to avoid signal aliasing at a distance z, the discrete version of the measurements for the three main zones can be written as

(a) Near zone:

$$y_{k,l}^{\ell} = \left| \sum_{s,r}^{M} \left(\sum_{k,l}^{M} \tilde{o}_{k,l}^{\ell} e^{2j\pi\left(\frac{ks}{M} + \frac{lr}{M}\right)} \right) T_{s,r} e^{-2j\pi\left(\frac{ks}{M} + \frac{lr}{M}\right)} \right|^2, \tag{30}$$

where $T_{s,r} = e^{\frac{-j2\pi z}{\lambda}\sqrt{1 - \frac{\lambda^2(s^2+r^2)}{M^2\Delta_s^2}}}$, is the discrete version of $T(k_x,k_y;z)$.

(b) Middle zone:

$$y_{\tilde{k},\tilde{l}}^{\ell} = \left| c_z^{\lambda} e^{\frac{j\pi\Delta_s^2}{z\lambda}(\tilde{k}^2+\tilde{l}^2)} \sum_{k,l}^{S} \tilde{o}_{k,l}^{\ell} e^{\frac{j\pi z\lambda}{(S\Delta_s)^2}(k^2+l^2)} e^{-2j\pi\left(\frac{\tilde{k}k}{S} + \frac{\tilde{l}l}{S}\right)} \right|^2 \tag{31}$$

(c) Far zone:

$$y_{s,r}^{\ell} = \frac{\Delta_{\tilde{o}}^4}{\lambda^2 z^2} \left| \sum_{k,l}^{M} \left(\sum_{s=kr_m}^{(k+1)r_m} \sum_{r=lr_m}^{(l+1)r_m} M_{s,r}^{\ell} o_{s,r} \right) e^{2j\pi\left(\frac{ks}{M} + \frac{lr}{M}\right)} \right|^2. \tag{32}$$

3.1.1. ***General Forward Matrix Model .*** From the discrete model previously developed, a matrix representation of the diffraction process is presented. Let $\mathbf{X}, \mathbf{M}_{\ell} \in \mathbb{C}^{N\times N}$ with $N = r_m S$ and $\mathbf{Q}_1 \in \mathbb{C}^{S\times S}$ be matrices whose entries are given by

$$\begin{aligned}
\mathbf{X}[t,q] &= o_{t,q}, \\
\mathbf{M}_\ell[t,q] &= M^\ell_{t,q}, \\
\mathbf{Q}_1[k,l] &= e^{\frac{j\pi z\lambda}{(M\Delta_s)^2}(k^2+l^2)}. \\
\mathbf{T}[x,y] &= e^{\frac{-j2\pi z}{\lambda}\sqrt{1-\frac{\lambda^2(s^2+r^2)}{M^2\Delta_s}}}
\end{aligned} \tag{33}$$

Considering (33), an expression in vector form of measurements in each zone can be obtained. Specifically, taking $\mathbf{x} \in \mathbb{C}^n$ as a column-wise vectorization of the matrix $\mathbf{X}$, and the diagonal matrices $\tilde{\mathbf{M}}_\ell \in \mathbb{C}^{n\times n}$ and $\tilde{\mathbf{Q}} \in \mathbb{C}^{m\times m}$, whose elements are the entries of the matrix $\mathbf{M}_\ell$ and $\mathbf{Q}$, respectively, the discrete versions can be modeled as

$$\begin{aligned}
\mathbf{y}_\ell &= |\mathbf{FTF}^H\mathbf{DM}_\ell\mathbf{x}|^2 && (\textit{Near zone}) , \\
\mathbf{y}_\ell &= |\mathbf{F}^H\mathbf{QDM}_\ell\mathbf{x}|^2 && (\textit{Middle zone}) , \\
\mathbf{y}_\ell &= |\mathbf{FDM}_\ell\mathbf{x}|^2 && (\textit{Far zone}) ,
\end{aligned} \tag{34}$$

where $n = N^2$, $m = S^2$, $\mathbf{y}_\ell \in \mathbb{R}^m_+$, and $\mathbf{D} \in \mathbb{R}^{m\times n}$ represents a down-sampling matrix defined as

$$(\mathbf{D})_{i,b} = \begin{cases} \frac{1}{r_m^2}, & \text{if} \quad \tilde{i} = \lfloor \frac{\tilde{b}(\mathrm{mod}N)}{r_m} \rfloor + 1 \ \textit{and} \\ & \quad (\tilde{i}-i)r_m^2 = (\tilde{b}-b). \\ 0, & \text{otherwise}, \end{cases} \tag{35}$$

where $\tilde{i} = i - \frac{N}{r_m}\lfloor\frac{(i-1)r_m}{N}\rfloor$ and $\tilde{b} = b - Nr_m\lfloor\frac{(b-1)}{Nr_m}\rfloor$, such that $\mathbf{D}\tilde{\mathbf{M}}_\ell\mathbf{x}$ models (29), and $\mathbf{F} \in \mathbb{C}^{m\times m}$ is the discrete Fourier transform matrix with entries

$$\mathbf{f}_i^H = \frac{1}{\sqrt{m}}[\omega^{-0(i-1)}, \omega^{-1(i-1)}, \cdots, \omega^{-(n-1)(i-1)}], \tag{36}$$

such that $i = 1, \cdots, m$ and $\omega = e^{\frac{2\pi j}{m}}$ is the m-th root of unity. Moreover, if $\mathbf{g}_2 = [\mathbf{y}_1, \cdots, \mathbf{y}_L]$ is defined as the global measurement vector at the middle zone, we have

$$\mathbf{g}_2 = |\mathbf{A}_2\mathbf{x}|^2, \tag{37}$$

where the matrix $\mathbf{A}_2$ is the vertical concatenation of the matrices $\mathbf{F}\tilde{\mathbf{Q}}_1\mathbf{D}\tilde{\mathbf{M}}_\ell$ given by

$$\mathbf{A}_2 = \frac{\Delta_o^2}{\lambda z}\left[(\mathbf{F}\tilde{\mathbf{Q}}_1\mathbf{D}\tilde{\mathbf{M}}_1)^H, \cdots, (\mathbf{F}\tilde{\mathbf{Q}}_1\mathbf{D}\tilde{\mathbf{M}}_L)^H\right]^H. \tag{38}$$

Additionally, the low-resolution sensing models for the other diffraction zones are summarized in Table 1. Specifically, for the near zone the matrix $\mathbf{A}_1$ is defined as

$$\mathbf{A}_1 = \left[(\mathbf{F}\mathbf{T}\mathbf{F}^H\mathbf{D}\tilde{\mathbf{M}}_1)^H, \cdots, (\mathbf{F}\mathbf{T}\mathbf{F}^H\mathbf{D}\tilde{\mathbf{M}}_L)^H\right]^H, \tag{39}$$

where $\mathbf{T}$ represents the discrete spatial frequency transfer function. Similarly, the sensing matrix $\mathbf{A}_3$ corresponding to the far field can be modeled as

Table 1. State-of-the-art Super-resolution discrete Models

Discrete Model	Near zone	Middle zone	Far zone
Proposed	$\mathbf{g}_1 = \lvert\mathbf{A}_1\mathbf{x}\rvert^2$	$\mathbf{g}_2 = \lvert\mathbf{A}_2\mathbf{x}\rvert^2$	$\mathbf{g}_3 = \lvert\mathbf{A}_3\mathbf{x}\rvert^2$
Katkovnik et al. (2017)	$\mathbf{D}\lvert\mathbf{G}_1\mathbf{x}\rvert^2$	-	-
Katkovnik and Egiazarian (2017)	-	-	$\mathbf{D}\lvert\mathbf{G}_2\mathbf{x}\rvert^2$
Jaganathan et al. (2016)	-	-	$\lvert\mathbf{G}_3\mathbf{x}\rvert^2$

$$\mathbf{A}_3 = \frac{\Delta_o^2}{\lambda z}\left[(\mathbf{FD}\tilde{\mathbf{M}}_1)^H, \cdots, (\mathbf{FD}\tilde{\mathbf{M}}_L)^H\right]^H. \tag{40}$$

To compare the derived matrix models in (38), (39) and (40) with those from the state-of-the-art. Table 1 summarizes the different super-resolution phase retrieval models for coded diffraction patterns. In particular, the matrices $\mathbf{G}_1$ and $\mathbf{G}_2$ are given by $\mathbf{G}_1 = \left[(\mathbf{FTF}^H\tilde{\mathbf{M}}_1)^H, \cdots, (\mathbf{FTF}^H\tilde{\mathbf{M}}_L)^H\right]$ and $\mathbf{G}_2 = \left[(\mathbf{F}^H\tilde{\mathbf{M}}_1)^H, \cdots, (\mathbf{F}^H\tilde{\mathbf{M}}_L)^H\right]$ from Katkovnik et al. (2017); Katkovnik and Egiazarian (2017), respectively. On the other hand, Jaganathan et al. (2016) assumes that the measurements are low-frequencies, i.e. matrix $\mathbf{G}_3$ is modeled as $\mathbf{G}_3 = \left[(\mathbf{SF}^H\tilde{\mathbf{M}}_1)^H, \cdots, (\mathbf{SF}^H\tilde{\mathbf{M}}_L)^H\right]$, where $\mathbf{S}$ is a diagonal selection matrix that only chooses the low frequencies. Under the derived models in (38), (39), and (40), the next section provides uniqueness guarantees for the super-resolution phase retrieval problem from CDP.

3.2. Uniqueness guarantees for physical super resolution phase retrieval

From the vector form of the proposed super-resolution models in Table 1, each measurement $(\mathbf{g}_u)_i$ is modeled as

$$(\mathbf{g}_u)_i = \lvert\mathbf{a}_{u,i}^H\mathbf{x}\rvert^2 = \mathbf{a}_{u,i}^H\mathbf{x}\mathbf{x}^H\mathbf{a}_{u,i}, \tag{41}$$

where $\mathbf{a}_{u,i}$ is the i-th row of the matrix $\mathbf{A}_u$, for the u-th diffraction zone, with $u = 1,2,3$. Let $\mathscr{A}_u : \mathscr{S}_+^{n\times n} \to \mathbb{R}^{mL}$ be the linear mapping defined as

$$\mathscr{A}_u(\mathbf{W}) = [\mathbf{a}_{u,1}^H \mathbf{W}\mathbf{a}_{u,1}, \cdots, \mathbf{a}_{u,mL}^H \mathbf{W}\mathbf{a}_{u,mL}]^T, \tag{42}$$

where $\mathbf{W}$ is a matrix variable that belongs to $\mathscr{S}_+^{n\times n}$ which is the space of self-adjoint positive semidefinite matrices. Observe that if $\mathbf{W} = \mathbf{x}\mathbf{x}^H$ then $\mathbf{g}_u = \mathscr{A}_u(\mathbf{W})$ for all $u = 1,2,3$. In order to guarantee unique solution from the phaseless measurements, the linear operators $\mathscr{A}_u(\cdot)$ must be injective Candes et al. (2013); Gross et al. (2017). Indeed, it is only necessary to prove that $\mathscr{A}_u(\cdot)$ is injective in the set

$$\mathscr{T}_{\mathbf{x}} := \{\mathbf{x}\mathbf{w}^H + \mathbf{w}\mathbf{x}^H : \mathbf{w} \in \mathbb{C}^n\}, \tag{43}$$

which is the tangent space of the manifold of all rank-one Hermitian matrices at the point $\mathbf{x}\mathbf{x}^H$ Candes et al. (2013); Gross et al. (2017). Proving this property over $\mathscr{A}_u(\cdot)$ guarantees the existence of a unique solution to the proposed super-resolution phase retrieval scenario Candes et al. (2013); Gross et al. (2017). Before proving Theorem 1, it is important to remark that this thesis assumes that the coded aperture $\tilde{\mathbf{M}}_\ell$ entries are $i.i.d$ copies of an admissible random variable d which satisfies Definition 1.

Definition 1. (*Admissible Random Variable*). A discrete random variable obeying $|d| \leq 1$, is said to be *admissible*.

Theorem 1. Fix any $\delta \in (0,1)$ and the set of coded apertures $\{\tilde{\mathbf{M}}_\ell : \ell = 1,\cdots,L\}$ with *i.i.d* entries of an admissible random variable d. Assume that for some constant $c > 0$, the matrix $\mathbf{P} = \sum_{\ell=1}^{L} \tilde{\mathbf{M}}_\ell^H \mathbf{D}^H \mathbf{D} \tilde{\mathbf{M}}_\ell$ satisfies

$$\|\mathbf{P} - c\mathbf{I}\|_\infty^2 \leq \delta, \tag{44}$$

where $L \geq c_0 n$, for some sufficiently large constant $c_0 > 0$, with $\mathbf{I}$ as the identity matrix, and $\mathbf{D}$ the down-sampling matrix as in (35). Then, considering (44), the sensing matrices $\mathbf{A}_u$ for the u-th diffraction zone satisfy

$$\mathscr{P}\left(\frac{1}{cmL}\|\mathbf{A}_u\|_\infty^2 \leq 1+\delta\right) \leq 1 - ne^{-c_1 mL\varepsilon^2}, \tag{45}$$

for some constant $c_1 > 0$ with $\varepsilon := \max(\delta, \delta^2)$. Also, with the same probability of (45), the linear operators $\mathscr{A}_u(\cdot)$ that model the phaseless measurements for the u-th diffraction zone are injective in the set $\mathscr{T}_{\mathbf{x}}$, that is

$$(1-\delta)\|\mathbf{W}\|_1 \leq \frac{1}{cmL}\|\mathscr{A}_u(\mathbf{W})\|_1 \leq (1+\delta)\|\mathbf{W}\|_1, \tag{46}$$

for any $\mathbf{W} \in \mathscr{T}_{\mathbf{x}}$. Therefore, the existence of a unique solution to the proposed super-resolution phase retrieval scenario is guaranteed with high probability.

Demostración. Let $\mathbf{W} \in \mathscr{T}_{\mathbf{x}}$, have rank at most two. For a normalized eigenvector of $\mathbf{W}$, the eigen-

value decomposition can be expressed as

$$\mathbf{W} = \lambda_1 \mathbf{b}\mathbf{b}^H + \lambda_2 \mathbf{v}\mathbf{v}^H, \tag{47}$$

with non-negative eigenvalues λ_1, λ_2 and normalized vectors $\mathbf{b}, \mathbf{v}$. Observe that from the definition of the linear map $\mathscr{A}_u(\cdot)$ in (42) we have that

$$\begin{aligned} \|\mathscr{A}_u(\mathbf{W})\|_1 &= \sum_{i=1}^{mL} \left| \lambda_1 |\mathbf{a}_{i,u}\mathbf{b}|^2 + \lambda_2 |\mathbf{a}_{i,u}\mathbf{v}|^2 \right| \\ &\leq \sum_{i=1}^{mL} |\lambda_1| \, |\mathbf{a}_{i,u}\mathbf{b}|^2 + |\lambda_2| \, |\mathbf{a}_{i,u}\mathbf{v}|^2 \\ &= |\lambda_1| \|\mathbf{A}_u\mathbf{b}\|_2^2 + |\lambda_2| \|\mathbf{A}_u\mathbf{v}\|_2^2 \\ &\leq (|\lambda_1| + |\lambda_2|) \|\mathbf{A}_u\|_\infty^2 = \|\mathbf{W}\|_1 \|\mathbf{A}_u\|_\infty^2, \end{aligned} \tag{48}$$

in which the first and second inequalities are obtained using the triangle inequality, and the last claim is based on the fact that $\sum_j |\lambda_j| = \|\mathbf{W}\|_1$ and $\|\mathbf{b}\|_2 = \|\mathbf{v}\|_2 = 1$. On the other hand, notice that from the definition of $\mathbf{A}_u$ in Table 1, it can be obtained that

$$\begin{aligned} \|\mathbf{A}_u\|_\infty^2 &= \lambda_{max}\left(\mathbf{A}_u^H \mathbf{A}_u\right) \\ &= \left\| \sum_{\ell=1}^{L} \tilde{\mathbf{M}}_\ell^H \mathbf{D}^H \mathbf{D} \tilde{\mathbf{M}}_\ell \right\|_\infty = \|\mathbf{P}\|_\infty^2, \end{aligned} \tag{49}$$

where $\lambda_{max}(\cdot)$ denotes the largest eigenvalue of a matrix. Additionally, assuming the condition in (44) holds for $L \geq c_0 n$, for some sufficiently large constant $c_0 > 0$, then from Theorem 5.44 in

Vershynin (2010) it can be obtained that

$$\mathscr{P}\left(\frac{1}{cmL}\|\mathbf{A}_u\|_\infty^2 \leq 1+\delta\right) \leq 1-ne^{-c_1 mL\varepsilon^2}, \tag{50}$$

for some constant $c_1 > 0$ with $\varepsilon := \max(\delta, \delta^2)$. Thus, combining (48) and (50), the right side of the inequality in (46) is expressed as

$$\frac{1}{cmL}\|\mathscr{A}_u(\mathbf{W})\|_1 \leq (1+\delta)\|\mathbf{W}\|_1 \tag{51}$$

with probability at least $1-ne^{-c_1 mL\varepsilon^2}$.

On the other hand, since (50) holds, from Lemma 5.36 in Vershynin (2010) it can be obtained that

$$\begin{aligned}\frac{1}{c}\|\mathscr{A}_u(\mathbf{W})\|_1 &= \frac{1}{c}\left(\lambda_1\|\mathbf{A}_u\mathbf{b}\|_2^2 + \lambda_2\|\mathbf{A}_u\mathbf{v}\|_2^2\right)\\ &\geq (1-\delta)(\lambda_1+\lambda_2) = (1-\delta)\|\mathbf{W}\|_1,\end{aligned} \tag{52}$$

with probability at least $1-ne^{-c_1 mL\varepsilon^2}$, where the last equality is obtained since $\mathbf{W}$ is positive semidefinite. Thus, from (52)

$$\frac{1}{cmL}\|\mathscr{A}_u(\mathbf{W})\|_1 \geq \frac{1}{mL}(1-\delta)\|\mathbf{W}\|_1, \tag{53}$$

Finally, combining the left side of (51) and the right side of (53), the result in (46) holds. In consequence, each operator $\mathscr{A}_u(\cdot)$ is injective with high probability, guaranteeing the existence of a unique solution to the proposed super-resolution phase retrieval scenario Candes et al. (2015a);

Candès and Li (2014). □

Notice that Theorem 1 proves that the probability of the linear operators $\mathscr{A}_u(\cdot)$ to be injective increases when the value of δ in (44) is small. From an optimization point of view, this result implies the need to design the set of coded apertures to guarantee uniqueness. Then, considering this observation, the following section provides a strategy to design the set of coded apertures based on condition (44).

4. Coded Aperture Design in Phase Retrieval

As shown in the theoretical and experimental results, the coded apertures (CA) distribution plays a crucial role in recovering the phase in coded diffraction patterns (CDP). Therefore, this chapter presents two CA design strategies, one independent of the data, based on a greedy strategy to increase the theoretical recovery probability. The other is based on data, using an end-to-end (E2E) deep learning approach.

4.1. Greedy strategy based on uniqueness guarantees

The result from the previous section provides useful guidelines for the design of coded-aperture ensembles, which lead to satisfy (46) so that better sensing matrices can be obtained to guarantee uniqueness. In particular, the theoretical condition in (44) shows that the set of coded apertures defines the concentration of measure of the largest eigenvalue of the sensing matrix $\mathbf{A}_u$ Hinojosa et al. (2018). Then, to determine a strategy to design the set of coded apertures, the structure of matrix $\mathbf{P}$ has to be analyzed. Thus, notice that from (35) we have that

$$\mathbf{D}^H\mathbf{D} = \frac{1}{r_m^2}\mathbf{I} + \mathbf{R}, \tag{54}$$

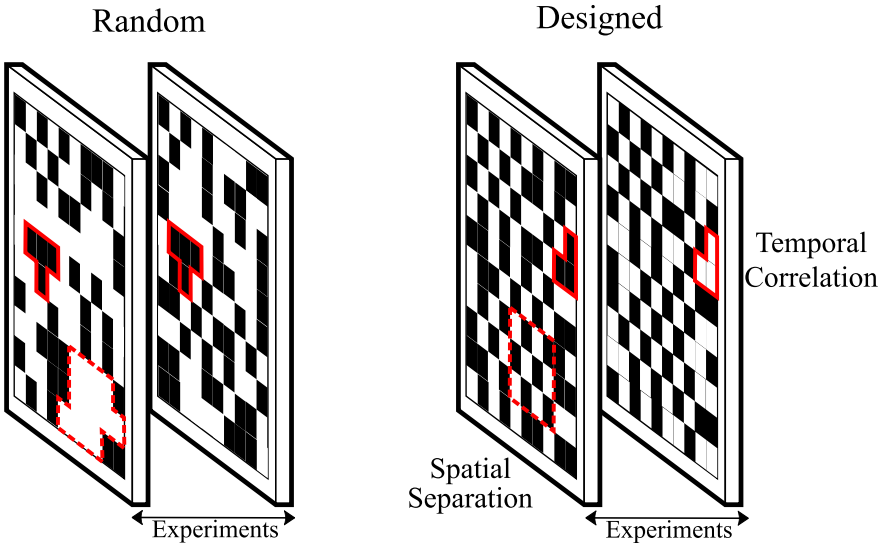
where $\mathbf{R}$ contains the off-diagonal terms of $\mathbf{D}^H\mathbf{D}$. Taking (54) into account, the matrix $\mathbf{P}$ can be expressed as

$$\mathbf{P} = \sum_{\ell=1}^{L} \tilde{\mathbf{M}}_\ell^H \mathbf{D}^H \mathbf{D} \tilde{\mathbf{M}}_\ell = \underbrace{\frac{1}{r_m^2} \sum_{\ell=1}^{L} \tilde{\mathbf{M}}_\ell^H \tilde{\mathbf{M}}_\ell}_{\mathbf{V}_1} + \underbrace{\sum_{\ell=1}^{L} \tilde{\mathbf{M}}_\ell^H \mathbf{R} \tilde{\mathbf{M}}_\ell}_{\mathbf{V}_2}. \tag{55}$$

From (55), it can be observed that (44) can be satisfied if $\mathbf{V}_1 = c\mathbf{I}$, and $\mathbf{V}_2 = \mathbf{0}$ for some imposed constant $c > 0$, where $\mathbf{0}$ represents a $n \times n$ all-zero matrix. More precisely, these two conditions lead to the following design criteria:

(a) Temporal correlation: Condition $\mathbf{V}_1 = c\mathbf{I}$ for some $c > 0$ can be accomplished if each pixel of the image is modulated for all the coding elements of the random variable d, along the L-experiments.

Figure 7. Visual comparison between a designed and a non-designed coded aperture.



Note: For two experiments with the admissible random variable $d \in \{1, 0\}$.

(b) Spatial separation: In practical terms, one can minimize the term $\mathbf{V}_2$ by building a set of coded apertures with an equispaced distribution of the coding elements, since $\mathbf{R}$ is the matrix that contains the off-diagonal terms of $\mathbf{D}^H\mathbf{D}$.

A recent work has developed a strategy that considers these criteria to design the coded apertures Mejia and Arguello (2018). Specifically, Mejia and Arguello (2018) minimizes the upper bounds of the Gershgorin theorem of a given matrix, which in this case is $\mathbf{P}$. This process generates a uniform distribution of the coding elements within the coded apertures ensuring that $\mathbf{V}_1 = c\mathbf{I}$, for some $c > 0$ and $\mathbf{V}_2 \approx \mathbf{0}$. This thesis follows the optimization strategy developed in Mejia and Arguello (2018) to design the set of coded apertures, which is formulated as

$$\begin{aligned}\min_{\{\tilde{\mathbf{M}}_\ell\}} \quad & \left\| \mathbf{1}_n^T \sum_{\ell=1}^{L} (\tilde{\mathbf{M}}_\ell^H \mathbf{D}^H \mathbf{D} \tilde{\mathbf{M}}_\ell) - (U/n)\mathbf{1}_n^T \right\|_2^2 \\ & + \left\| \sum_{\ell=1}^{L} (\tilde{\mathbf{M}}_\ell^H \mathbf{D}^H \mathbf{D} \tilde{\mathbf{M}}_\ell)\mathbf{1}_n - (U/n)\mathbf{1}_n \right\|_2^2, \end{aligned} \tag{56}$$

for $\ell = 1, \cdots, L$, where $\mathbf{1}_n \in \mathbb{R}^n$ denotes the vector whose entries are ones, and U is a constant. This optimization problem is solved using a greedy algorithm Mejia and Arguello (2018). The design criteria in (56) can be referred to as uniform sensing, as illustrated in Fig. 7, for an admissible random variable $d \in \{1, 0\}$. Specifically, the first term in (56) handles the sum per column of $\mathbf{P}$ that indicates the number of times a pixel of the image is sensed. Additionally, the second term of (56) indicates the number of pixels of the image measured onto a certain experiment by the sensor. In addition, the proposed coded aperture design prevents forming clusters of the same coding element, as happens in Fig. 7. It is worth mentioning that Fig. 7 illustrates a designed coded aperture when

d has just two possible values i.e. $\{0,1\}$. In the case when d has more than two values, the coded aperture can be seen in Mejia and Arguello (2018) as a three-dimensional binary structure, where the third dimension is related to the number of elements of d, such that each slice corresponds to a binary mask that indicates whether each value of d is located.

4.2. End-to-End Phase Mask Design

This proposed strategy seeks the sensing design in PR based on MPM for reconstruction procedure using the E2E approach. Here, it jointly optimizes the MPM $\mathbf{D} \in \{\mathbf{D}_\ell\}_{\ell=1}^{L}$, and the parameters θ, according to a chosen DNN $\mathscr{M}_\theta(\cdot)$, by minimizing a cost function $\mathscr{L}(\cdot)$ considering a regularization function $\mathscr{R}(\cdot)$ that promotes particular properties in MPM. From a set of $\mathscr{J}$ scenes $\{\mathbf{x}^{(j)}\}_{j=1}^{\mathscr{J}}$, which produce the initial estimation $\{\tilde{\mathbf{z}}^{(j)} = \text{FSI}(|\mathbf{f}_k^H \mathbf{D}_\ell \mathbf{x}^{(j)}|^2)\}_{j=1}^{\mathscr{J}}$, then, the proposed optimization problem is given by

$$\begin{aligned} \{\mathbf{D}^*, \theta^*\} &\in \arg\min_{\mathbf{D},\theta} \mathscr{L}(\mathbf{D}, \theta | \mathbf{x}^{(j)}, \tilde{\mathbf{z}}^{(j)}), \\ \mathscr{L}(\mathbf{D}, \theta | \mathbf{x}^{(j)}, \mathbf{z}^{(j)}) &= \frac{1}{\mathscr{J}} \sum_{j=1}^{\mathscr{J}} ||\mathscr{M}_\theta(\tilde{\mathbf{z}}^{(j)}) - \mathbf{x}^{(j)}||_2^2 + \rho \mathscr{R}(\mathbf{D}), \end{aligned} \tag{57}$$

where $\rho > 0$ is a regularization parameter. Algorithm 2 summarizes the proposed E2E methodology in order to solve the optimization problem stated in (57). In line 5, the acquisition model is implemented. Then, in line 6, the optical field is estimated through Algorithm 1. The designed DNN refines the optical field approximation in line 7. The loss function is evaluated in line 8. Besides, the gradients of $\mathbf{D}$ and θ are computed in lines 9 and 10, respectively, which are used in the Adam update Kingma and Ba (2014). Finally, the optimal MPM and the optimal parameters θ of

the recovery phase network are returned in line 11.

Algorithm 2 Learning MPM.

1: **Input:** Training set $\{\mathbf{x}^{(j)}\}_{j=1}^{\mathscr{J}}$ with $\mathscr{J}$ images.
2: **Initialize:** Initialize the optimization variables for L MPM as $\mathbf{D} \in \{\mathbf{D}_\ell\}_{\ell=1}^{L}$ from a uniform distribution.
3: **for** epoch = 1:$\mathscr{E}$ **do**
4: **for** $j = 1$: $\mathscr{J}$ **do**
5: $y_{k,\ell}^{(j)} = |\mathbf{f}_k^H \mathbf{D}_\ell \mathbf{x}^{(j)}|^2$, $k \in \{1, \cdots, n\}$
6: $\tilde{\mathbf{z}}^{(j)} \leftarrow \text{FSI}\left(y_{k,\ell}^{(j)}, \mathbf{D}_\ell\right)$
7: $\mathbf{z}^{(j)} \leftarrow \mathscr{M}_\theta\left(\tilde{\mathbf{z}}^{(j)}\right)$
8: $\mathscr{L}_{\mathbf{D},\theta} = \frac{1}{J}\sum_{j=1}^{\mathscr{J}} \|\mathbf{x}^{(j)} - \mathbf{z}^{(j)}\|_2^2 + \rho\mathscr{R}(\mathbf{D})$
9: $\mathbf{D} \leftarrow \mathscr{A}_{dam}(\mathbf{D}, \beta_1 \nabla_{\mathbf{D}} \mathscr{L}_{\mathbf{D},\theta})$
10: $\theta \leftarrow \mathscr{A}_{dam}(\theta, \beta_2 \nabla_\theta \mathscr{L}_{\mathbf{D},\theta})$
11: **end for**
12: **end for**
13: **Return:** Optimal MPM $\mathbf{D}$ and θ

5. Proposed Phase Retrieval Recovery Methods

Recent literature has shown that the non-convex formulations outperform the convex methods, requiring fewer measurements and less computational complexity to retrieve the phase image. However, most non-convex methods are based on non-smooth loss function, and they do not include prior information about the signal, such as sparsity. Therefore, this chapter presents some algorithms proposed based on a smoothed non-convex least-squares objective function, where sparsity prior as total-variation and deep priors are also included in the proposed formulation.

5.1. Smoothing Phase Retrieval Algorithm

The smoothing phase retrieval problem is proposed to the general phase retrieval problem formulated as the system of m quadratic equations of the form

$$y_k = |\langle \mathbf{a}_k, \mathbf{x} \rangle|^2, k = 1, \cdots, m, \tag{58}$$

where the data vector $\mathbf{y} := [y_1, \cdots, y_m]^T \in \mathbb{R}^m$ represents the measurements, $\mathbf{a}_k \in \mathbb{R}^n/\mathbb{C}^n$ are the known sampling vectors and $\mathbf{x} \in \mathbb{R}^n/\mathbb{C}^n$ is the desired unknown signal. This thesis considers the complex-valued Gaussian design vectors as $\mathbf{a}_k \sim \mathscr{CN}(0, \mathbf{I}_n) = \mathscr{N}(0, \frac{1}{2}\mathbf{I}_n) + j\mathscr{N}(0, \frac{1}{2}\mathbf{I}_n)$, assumed to be independently and identically distributed (i.i.d.), where $j = \sqrt{-1}$. For the real Gaussian case the sampling vectors $\mathbf{a}_k$ are given by $\mathbf{a}_k \sim \mathscr{N}(0, \mathbf{I}_n)$, also assumed to be i.i.d.. Then, adopting the least-squares criterion, the task of recovering a solution from the phaseless measurements in Eq.

(1) reduces to that of minimizing the amplitude-based lost function

$$\min_{\mathbf{x}\in\mathbb{C}^n} f(\mathbf{x}) = \min_{\mathbf{x}\in\mathbb{C}^n} \frac{1}{m}\sum_{k=1}^{m} |\langle \mathbf{a}_k, \mathbf{x}\rangle| - \sqrt{y_k})^2 . \tag{59}$$

Notice that the optimization problem in (59) is non-smooth and non-convex Pinilla et al. (2018a). Then, this method proposes an algorithm based on the Smoothing Projected Gradient (SPG) method Zhang and Chen (2009), to solve this non-smooth and non-convex optimization problem. This method uses an auxiliary smoothing function $g(\cdot)$ to approximate the original objective function, in order to solve the non-smooth and non-convex optimization problem.

Definition 2. *Smoothing function:* Let $f : \mathbb{C}^n \to \mathbb{R}$ be a locally Lipschitz continuous function. Then $g : \mathbb{C}^n \times \mathbb{R}_+ \to \mathbb{R}$ is a smoothing function of $f(\cdot)$, if $g(\cdot, \mu)$ is smooth in $\mathbb{C}^n$ for any fixed $\mu \in \mathbb{R}_{++}$ and

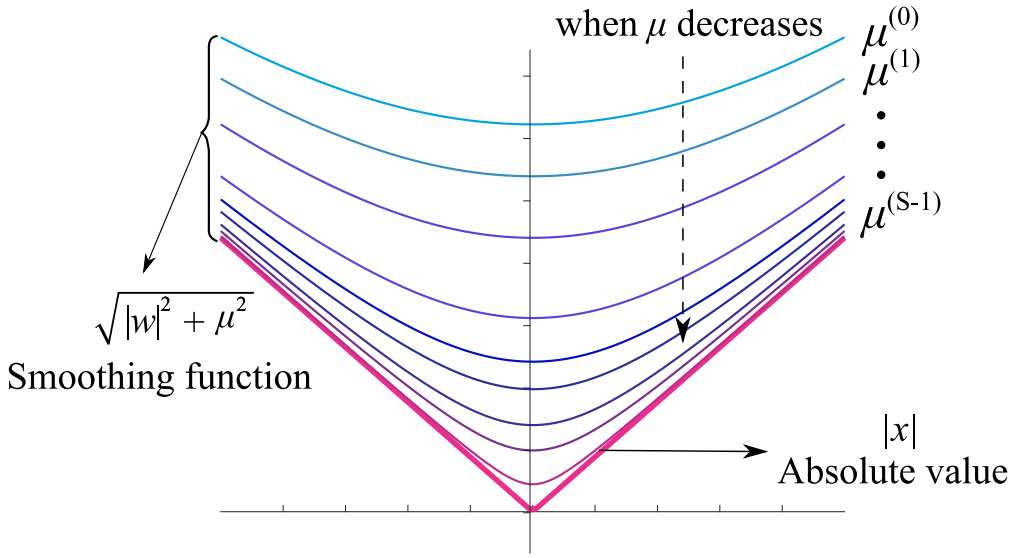
$$\lim_{\mu\downarrow 0} g(\mathbf{w}, \mu) = f(\mathbf{w}), \tag{60}$$

for any fixed $\mathbf{w} \in \mathbb{C}^n$.

Smoothing methods are widely used in non-smoothing functions such as TV regularization, sparsity, among others Zhang and Chen (2009). However, this technique has not yet been used for the phase retrieval problem. Therefore, according to the above definition, this thesis considers the function $\varphi_\mu : \mathbb{R} \to \mathbb{R}_{++}$ defined as

$$\varphi_\mu(w) = \sqrt{w^2 + \mu^2}, \tag{61}$$

Figure 8. Visual representation of the smoothing function.



Note: The smoothing function tending to the absolute value mapping when μ decreases.

where $\mu \in \mathbb{R}_{++}$ is a tunable parameter, that decreases at each iteration as illustrated in Fig. 8. The following lemma shows that $\varphi_\mu(\cdot)$ has important smooth properties to approximate the functions $f_k(\cdot)$, given that $\varphi_0(|\mathbf{a}_k^H \mathbf{x}|) = f_k(\mathbf{x})$.

Lemma 1. The function $\varphi_\mu(w)$, defined in Eq. (61), has the following properties.

1. $\varphi_\mu(w)$ is Lipschitz continuous function.

2. $\varphi_\mu(w)$ converges uniformly to $\varphi_0(w)$ on $\mathbb{R}$.

Demostración. 1. Since $\mu > 0$ then $\varphi_\mu(w)$ is smooth on $\mathbb{R}$, where $\varphi'_\mu(w)$ is given by

$$\varphi'_\mu(w) = \frac{w}{\sqrt{w^2+\mu^2}}. \tag{62}$$

Notice that $\sqrt{w^2+\mu^2} \geq w$ for all $w \in \mathbb{R}$, then $|\varphi'_\mu(w)| \leq 1$. Therefore, $\varphi_\mu(w)$ is a Lipschitz continuous function because its first derivative is bounded Eriksson et al. (2013). Further, the Lipschitz constant for the function $\varphi_\mu(\cdot)$ is $L_{\varphi_\mu} = 1$.

2. According to the definition of the function φ_μ in Eq. (61), it can be obtained that

$$|\varphi_\mu(w) - \varphi_0(w)| = |\sqrt{w^2+\mu^2} - \sqrt{w^2}|. \tag{63}$$

Note that by the Minkowski inequality Kreyszig (1989), it can be concluded that $\sqrt{w^2+\mu^2} \leq \sqrt{w^2}+\mu$, therefore

$$|\varphi_\mu(w) - \varphi_0(w)| \leq |\sqrt{w^2}+\mu-\sqrt{w^2}| \leq \mu. \tag{64}$$

□

The first result in Lemma 1 is used to guarantee the convergence of the proposed algorithm. Also, the second part of Lemma 1 establishes that the function $\varphi_\mu(|\mathbf{a}_k^H\mathbf{x}|)$ uniformly approximates $f_k(\mathbf{x})$, which is a desirable convergence, since it only depends on the value of μ. Therefore, a smooth optimization problem to recover the unknown desired signal $\mathbf{x} \in \mathbb{R}^n/\mathbb{C}^n$ from the measurements q_k in Eq. (59) can be formulated as

$$\min_{\mathbf{x}\in\mathbb{R}^n/\mathbb{C}^n} g(\mathbf{x},\mu) = \min_{\mathbf{x}\in\mathbb{R}^n/\mathbb{C}^n} \frac{1}{m}\sum_{k=1}^{m}\left(\varphi_\mu(|\mathbf{a}_k^H\mathbf{x}|) - q_k\right)^2, \tag{65}$$

where $q_k = \sqrt{y_k}$ and $g(\mathbf{x}, \mu)$ is the smoothing function of $f(\mathbf{x})$. Notice that if Eq. $\mu = 0$, (65) reduces to the non-smooth formulation described in (59). Theorem 2 shows that the function $g(\cdot)$ is a uniformly smooth approximation of the function $f(\cdot)$, which is a desired behavior in order to solve the optimization problem in Eq. (59).

Theorem 2. Let f and $g(\cdot, \mu)$ be as defined in Eq. (59) and Eq. (65), respectively. Then $g(\cdot, \mu)$ is smooth for any fixed $\mu > 0$, and there exists a constant $\kappa_1 > 0$ satisfying

$$|g(\mathbf{x}, \mu) - f(\mathbf{x})| \leq \mu \kappa_1. \tag{66}$$

Demostración. From Eqs. (59) and (65) it can be obtained that

$$|K(\mathbf{x}, \mu)| = \frac{1}{m} \left| \sum_{k=1}^{m} \left(\varphi_\mu(|\mathbf{a}_k^H \mathbf{x}|) - q_k\right)^2 - \left(\varphi_0(|\mathbf{a}_k^H \mathbf{x}|) - q_k\right)^2 \right|, \tag{67}$$

where $K(\mathbf{x}, \mu) = g(\mathbf{x}, \mu) - f(\mathbf{x})$. Note that the right hand side of the equality in Eq. (67) can be rewritten as

$$\frac{1}{m} \left| \sum_{k=1}^{m} \varphi_\mu^2(|\mathbf{a}_k^H \mathbf{x}|) - \varphi_0^2(|\mathbf{a}_k^H \mathbf{x}|) - 2q_k \left(\varphi_\mu(|\mathbf{a}_k^H \mathbf{x}|) - \varphi_0(|\mathbf{a}_k^H \mathbf{x}|)\right) \right|. \tag{68}$$

By definition of the function $\varphi_\mu(\cdot)$ in Eq. (61), it can be concluded that

$$\varphi_\mu^2(|\mathbf{a}_k^H \mathbf{x}|) - \varphi_0^2(|\mathbf{a}_k^H \mathbf{x}|) = \mu^2. \tag{69}$$

Applying the triangular inequality, it can be obtained that

$$|g(\mathbf{x}, \mu) - f(\mathbf{x})| \leq \frac{1}{m} \sum_{k=1}^{m} \mu^2 + 2q_k \left|\varphi_\mu(|\mathbf{a}_k^H \mathbf{x}|) - \varphi_0(|\mathbf{a}_k^H \mathbf{x}|)\right|. \tag{70}$$

Using the fact that the function $\varphi_\mu(\cdot)$ uniformly approximates the function $\varphi_0(\cdot)$ as was proved in Lemma 1, the above inequality can be expressed as

$$|g(\mathbf{x},\mu) - f(\mathbf{x})| \leq \frac{1}{m}\sum_{k=1}^{m} \mu^2 + 2q_k\mu. \tag{71}$$

Therefore by taking $q^{max} = \max\{q_k | k = 1, \cdots, m\}$, from Eq. (71) it can be obtained that

$$|g(\mathbf{x},\mu) - f(\mathbf{x})| \leq \frac{1}{m}\left(\sum_{k=1}^{m} \mu^2 + 2\mu q^{max}\right) = \mu\kappa_1, \tag{72}$$

where $\kappa_1 = (\mu + 2q^{max})$. Thus, the result holds.

On the other hand, the function g in Eq. (65) is smooth since $\varphi_\mu(|\mathbf{a}_k^H\mathbf{x}|)$ is smooth as defined in Eq. (61). □

On the other hand, to solve Eq. (65), this thesis presents the Phase Retrieval Smoothing Conjugate Gradient method (PR-SCG), summarized in Algorithm 3. This algorithm is a gradient descent method based on the SPG method. The PR-SCG uses a nonlinear conjugate gradient method developed in Chen and Zhou (2010), to accelerate its convergence. Following the algorithm in each iteration, a backtracking line search strategy is used to choose a correct step size of the conjugate gradient update direction, which is calculated in Line 9. Further, the smoothing parameter μ is updated as in Zhang and Chen (2009), to obtain a new point. That is, if $\|\partial g(\mathbf{x}_{i+1}, \mu_i)\|_2 \geq \gamma\mu_i$ in Line 10 is not satisfied, then the smoothing parameter is updated using the new point in Line 13. Algorithm 3 calculates the conjugate direction in Line 17. Each vector $\tilde{\mathbf{g}}_i$ in Algorithm 3 is cal-

culated using the Wirtinger derivative as was introduced in Hunger (2007). The following lemma introduces the Wirtinger derivative of the function $g(\mathbf{x}, \mu)$.

Lemma 2. The Wirtinger derivative of a real-valued function $h(\mathbf{z}) : \mathbb{C}^n \to \mathbb{R}$ with complex-valued argument $\mathbf{z} \in \mathbb{C}^n$ is obtained for

$$2\frac{\partial h(\mathbf{z})}{\partial \mathbf{z}^*}. \tag{73}$$

The proof of this lemma can be found in Corollary 5.0.1 in Hunger (2007). It is important to remark that this Wirtinger derivation has been recently used by the state-of-the-art methods to solve the phase retrieval problem Candes et al. (2015d); Wang et al. (2016a); Chen and Candes (2015b).

For simplicity, this thesis denotes the Wirtinger derivative of any function $h(\mathbf{z})$ as $\partial h(\mathbf{z})$, *i.e.* $\partial h(\mathbf{z}) = 2\frac{\partial h(\mathbf{z})}{\partial \mathbf{z}^*}$. Then, considering the result in Lemma 2, the Wirtinger derivative of $g(\mathbf{x}, \mu)$ is given by

$$\partial g(\mathbf{x}_i, \mu_i) = \frac{2}{m}\sum_{k=1}^{m}(\varphi_{\mu_i}(|\mathbf{a}_k^H\mathbf{x}_i|) - q_k)\partial\varphi_{\mu_i}(|\mathbf{a}_k^H\mathbf{x}_i|), \tag{74}$$

where

$$\partial\varphi_{\mu_i}(|\mathbf{a}_k^H\mathbf{x}_i|) = \frac{\mathbf{a}_k^H\mathbf{x}_i}{\varphi_{\mu_i}(|\mathbf{a}_k^H\mathbf{x}_i|)}\mathbf{a}_k. \tag{75}$$

Notice that, in contrast to the gradient update steps for the TAF and TWF methods introduced in Wang et al. (2017a) and Chen and Candes (2015b) respectively, $\partial g(\mathbf{x}_i, \mu_i)$ in Eq. (93) is always continuous because $\varphi_\mu(|\mathbf{a}_k^H\mathbf{w}|) \neq 0$ for any $\mathbf{w} \in \mathbb{C}^n$. Therefore, the proposed PR-SCG method does not require truncation parameters.

Algorithm 3 PR-SCG: Smoothing Conjugate Gradient Phase Retrieval Method

1: **Input:** Data $\{(\mathbf{a}_k; q_k)\}_{k=1}^m$ and $\varepsilon_0 = 10^{-10}$. Choose constants $\delta_1 = 0{,}9, \delta_2 = 0{,}4, \gamma_1 = 0{,}5, \mu_0 = 5\times 10^4/m, \gamma = 0{,}01$ and T maximum number of iterations.
2: Initial point $\mathbf{x}_0 = \sqrt{\frac{\sum_{k=1}^m q_k^2}{m}}\tilde{\mathbf{z}}_0$. where $\tilde{\mathbf{z}}_0$ is the leading eigenvector of $\mathbf{Y}_0 := \frac{1}{|I_0|}\sum_{k\in I_0}\sqrt{q_k}\frac{\mathbf{a}_k\mathbf{a}_k^H}{\|\mathbf{a}_k\|_2^2}$.
3: Set $\mathbf{d}_0 = -\tilde{\mathbf{g}_0} = -\partial g(\mathbf{x}_0, \mu_0)$.
4:
5: **for** $i = 0 : T-1$ **do**
6: Compute the step-size α_i by backtracking
7: Set $\rho = 1$.
8: **while** $g(\mathbf{x}_i + \rho\mathbf{d}_i, \mu_i) > g(\mathbf{x}_i, \mu_i) + \delta_1\rho\mathscr{R}\left(\tilde{\mathbf{g}}_i^H\mathbf{d}_i\right)$ **do**
9: $\rho = \delta_2\rho$
10: **end while**
11: $\alpha_i = \rho$ and $\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i\mathbf{d}_i$
12: **if** $\|\partial g(\mathbf{x}_{i+1}, \mu_i)\|_2 \geq \gamma\mu_i$ **then**
13: $\mu_{i+1} = \mu_i$
14: **else**
15: $\mu_{i+1} = \gamma_1\mu_i$
16: **end if**
17: $\tilde{\mathbf{p}}_i = \tilde{\mathbf{g}}_{i+1} - \tilde{\mathbf{g}}_i$ and $\mathbf{s}_i = \mathbf{x}_{i+1} - \mathbf{x}_i$.
18: $\tilde{\mathbf{z}}_i = \tilde{\mathbf{p}}_i + \left(\varepsilon_0\|\tilde{\mathbf{g}}_{i+1}\|_2^2 + \max\{0, -\frac{\mathscr{R}\left(\mathbf{s}_i^H\tilde{\mathbf{p}}_i\right)}{\|\mathbf{s}_i\|_2^2}\}\right)\mathbf{s}_i$.
19:

$$\mathbf{d}_{i+1} = -\tilde{\mathbf{g}}_{i+1} + \mathscr{R}\left(\frac{\tilde{\mathbf{g}}_{i+1}^H\tilde{\mathbf{z}}_i}{\mathbf{d}_i^H\tilde{\mathbf{z}}_i} - \frac{2\|\tilde{\mathbf{z}}_i\|_2^2\tilde{\mathbf{g}}_{i+1}^H\mathbf{d}_i}{|\mathbf{d}_i^H\tilde{\mathbf{z}}_i|^2}\right)\mathbf{d}_i + \mathscr{R}\left(\frac{\tilde{\mathbf{g}}_{i+1}^H\mathbf{d}_i}{\mathbf{d}_i^H\tilde{\mathbf{z}}_i}\right)\tilde{\mathbf{z}}_i.$$

20: **end for**
21: **return:** $\mathbf{x}_T$
22: **Notation:** $\mathscr{R}(\cdot)$ represents the real part function.

Initialization Stage. This thesis uses the Weighted Maximal Correlation initialization proposed in Wang et al. (2017b) and described in Section 2.2.3.1. This procedure is calculated in Line 2 of Algorithm 3. It is worth highlighting that any proposed initialization can be employed to be then refined by the proposed smoothing steps.

5.2. Stochastic Smoothing Phase Retrieval Algorithm

When the sample size is large, a stochastic algorithm is preferred due to its fast convergence, and low computational complexity Zhang and Liang (2016); Wang et al. (2017a). In this section, this thesis develops a stochastic algorithm named Stochastic Smoothing Phase Retrieval (SSPR), which is summarized in Algorithm 4. This thesis shows that SSPR guarantees exact recovery with a linear convergence rate.

The SSPR algorithm solves the following optimization problem

$$\min_{\mathbf{x}\in\mathbb{C}^n} g_1(\mathbf{x},\mu) = \min_{\mathbf{x}\in\mathbb{C}^n} \mathbb{E}\left[\ell_{k_t}(\mathbf{x},\mu)\right], \tag{76}$$

where $\mathbb{E}[\cdot]$ is the expected value function, and $\ell_{k_t}(\mathbf{x},\mu) = \left(\varphi_\mu(|\mathbf{a}_{k_t}^H\mathbf{x}|) - q_{k_t}\right)^2$ is a component function of $g(\mathbf{x},\mu)$ in Eq. (65), for some index $k_t \in \{1,2,\cdots,m\}$ per iteration $t \geq 0$. Specifically, SSPR successively updates $\mathbf{x}_0$ using the following stochastic gradient iterations for all $t \geq 0$

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha\left(\mathbf{a}_{k_t}^H\mathbf{x}_t - q_{k_t}\frac{\mathbf{a}_{k_t}^H\mathbf{x}_t}{\sqrt{|\mathbf{a}_{k_t}^H\mathbf{x}_t|^2+\mu_t^2}}\right)\mathbf{a}_{k_t}, \tag{77}$$

where $\alpha \in (0,1)$ is a constant. The index k_t is sampled uniformly at random from $\{1,2,\cdots,m\}$.

The following lemma establishes how to calculate the Wirtinger derivative of $g_1(\mathbf{x},\mu)$, which is a helpful result to prove the global convergence of the SSPR algorithm in Theorem 4.

Algorithm 4 SSPR: Stochastic Smoothing Phase Retrieval algorithm

Input: Data $\{(\mathbf{a}_k; q_k)\}_{k=1}^m$ and choose constants $\alpha = 1{,}6/n$. Choose $\gamma_1 = 0{,}9, \mu_0 = 6 \times 10^4/m, \gamma = 0{,}01$ and $T = 500m$ maximum number of iterations.

Initial point $\mathbf{x}_0 = \sqrt{\frac{\sum_{k=1}^m q_k^2}{m}}\tilde{\mathbf{x}}_0$, where $\tilde{\mathbf{x}}_0$ is the leading eigenvector of $\mathbf{Y}_0 := \frac{1}{|I_0|}\sum_{k\in I_0}\sqrt{q_k}\frac{\mathbf{a}_k\mathbf{a}_k^H}{\|\mathbf{a}_k\|_2^2}$.

for $t = 0 : T-1$ **do**

Choose k_t uniformly at random from $\{1, 2, \cdots, m\}$

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha\left(\mathbf{a}_{k_t}^H\mathbf{x}_t - q_{k_t}\frac{\mathbf{a}_{k_t}^H\mathbf{x}_t}{\sqrt{|\mathbf{a}_{k_t}^H\mathbf{x}_t|^2+\mu_t^2}}\right)\mathbf{a}_{k_t}$$

if $\|\partial g_1(\mathbf{x}_{t+1}, \mu_t)\|_2 \geq \gamma\mu_t$ **then**

$\mu_{t+1} = \mu_t$

else

$\mu_{t+1} = \gamma_1\mu_t$

end if

end for

return: $\mathbf{x}_T$

Lemma 3. The Wirtinger derivative of $g_1(\mathbf{x}, \mu)$ is given by

$$\partial g_1(\mathbf{x}, \mu) = \mathbb{E}\left[\partial \ell_{k_t}(\mathbf{x}, \mu)\right]. \tag{78}$$

Demostración. From Eq. (76) it has that

$$\begin{aligned} g_1(\mathbf{x}, \mu) &= \mathbb{E}\left[\left(\varphi_\mu(|\mathbf{a}_{k_t}^H\mathbf{x}|) - q_{k_t}\right)^2\right] \\ &= \mathbb{E}\left[\varphi_\mu^2(|\mathbf{a}_{k_t}^H\mathbf{x}|)\right] - 2\mathbb{E}\left[q_{k_t}\varphi_\mu(|\mathbf{a}_{k_t}^H\mathbf{x}|)\right] + \mathbb{E}[q_{k_t}^2]. \end{aligned} \tag{79}$$

Since k_t is sampled uniformly at random from $\{1,2,\cdots,m\}$ then

$$g_1(\mathbf{x},\mu) = \frac{1}{m}\sum_{k=1}^{m}\left(\varphi_\mu(|\mathbf{a}_k^H\mathbf{x}|) - q_k\right)^2 = \frac{1}{m}\sum_{k=1}^{m}\ell_k(\mathbf{x},\mu), \tag{80}$$

where $\ell_k(\mathbf{x},\mu) = \left(\varphi_\mu(|\mathbf{a}_k^H\mathbf{x}|) - q_k\right)^2$. From Eq. (80) it can be obtained that

$$\partial g_1(\mathbf{x},\mu) = \frac{2}{m}\sum_{k=1}^{m}(\varphi_\mu(|\mathbf{a}_k^H\mathbf{x}|) - q_k)\partial\varphi_\mu(|\mathbf{a}_k^H\mathbf{x}|). \tag{81}$$

On the other hand, notice that

$$\begin{aligned}\mathbb{E}\left[\partial\ell_{k_t}(\mathbf{x},\mu)\right] &= \mathbb{E}\left[2(\varphi_\mu(|\mathbf{a}_{k_t}^H\mathbf{x}|) - q_{k_t})\partial\varphi_\mu(|\mathbf{a}_{k_t}^H\mathbf{x}|)\right] \\ &= \frac{2}{m}\sum_{k=1}^{m}(\varphi_\mu(|\mathbf{a}_k^H\mathbf{x}|) - q_k)\partial\varphi_\mu(|\mathbf{a}_k^H\mathbf{x}|).\end{aligned} \tag{82}$$

Combining Eqs. (81) and (82) we have that

$$\partial g_1(\mathbf{x},\mu) = \mathbb{E}\left[\partial\ell_{k_t}(\mathbf{x},\mu)\right]. \tag{83}$$

Thus, from the above equation the result holds. □

The local error contraction of the step in Line 4 in Algorithm 4 is characterized by the following theorem, for any value of $\mu \in \mathbb{R}_{++}$.

Theorem 3. *Local error contraction:* Consider the noiseless measurements $q_k = |\langle\mathbf{a}_k,\mathbf{x}\rangle|$ for an arbitrary signal $\mathbf{x} \in \mathbb{C}^n$, and i.i.d $\{\mathbf{a}_k \sim \mathscr{CN}(0,\mathbf{I}_n)\}_{k=1}^m$. If $\alpha \in (0,\alpha_0/n]$ and also $m \geq c_0 n$ then,

with probability at least $1-2e^{-\varepsilon^2 m/2}$, the stochastic smoothing phase retrieval algorithm, tabulated in Algorithm 4, satisfies

$$\mathbb{E}_{k_t}\left[d_r^2(\mathbf{x}_{t+1},\mathbf{x})\right] \leq \rho\,(1-\upsilon)^{t+1}\,\|\mathbf{x}\|_2^2, \tag{84}$$

for $\rho = 1/10$ and some numerical constant $\upsilon \in (0,1)$, where the expectation is taken over the random variable k_t, and c_0 is a universal constant.

Demostración. The proof of Theorem 3 is relegated to Appendix A. □

Now, the following theorem, which uses the result in Theorem 3, establishes that the sequence $\{\mathbf{x}_i\}$ generated by Algorithm 4 reconstructs the solution exactly, up to a global uni-modular constant.

Theorem 4. In the setup of Theorem 3 we have that the sequences $\{\mu_i\}$ and $\{\mathbf{x}_i\}$ generated by Algorithm 4 satisfies

$$\lim_{i\to\infty}\mu_i = 0, \text{ and } \lim_{i\to\infty}\|\partial g_1(\mathbf{x}_i,\mu_{i-1})\|_2 = 0. \tag{85}$$

Demostración. The proof of Theorem 4 can be found in Appendix B. □

Notice that, combining Theorems 3 and 4, it can be concluded that the proposed SSPR algorithm achieves linear convergence because the number of equations m and unknowns exceeds a fixed numerical constant Chen and Candes (2015b).

5.3. Sparse Smoothing Phase Retrieval Algorithm

The sparse phase retrieval problem can be formulated as the solution to the system of m quadratic equations of the form

$$y_i = |\langle \mathbf{a}_i, \mathbf{x} \rangle|^2, i = 1, \cdots, m, \text{ subject to } \|\mathbf{x}\|_0 = k, \tag{86}$$

where $\mathbf{x} \in \mathbb{R}^n / \mathbb{C}^n$ is the desired unknown signal, the sparsity level $k \ll n$ is assumed to be known and $\|\cdot\|_0$ is the zero pseudo-norm. Notice that this prior information is incorporated in the traditional phase retrieval formulation. Similarly, this thesis adopts the smoothing formulation as

$$\min_{\|\mathbf{x}\|_0 = k} g(\mathbf{x}, \mu) = \frac{1}{m} \sum_{i=1}^{m} \left(\varphi_\mu(|\mathbf{a}_i^H \mathbf{x}|) - q_i \right)^2, \tag{87}$$

where $g(\mathbf{x}, \mu)$ is the smoothing function of $f(\mathbf{x})$ in (59).

To solve (87), this thesis proposes the Sparse Phase Retrieval algorithm via Smoothing Function (SPRSF), summarized in Algorithm 5. SPRSF is a gradient threshold descent method, which iteratively refines a initial guess solution. Specifically, in Line 2 the algorithm calculates the initial guess $\mathbf{z}^{(0)}$. Also, following the algorithm in each iteration, the threshold step is calculated in Line 4 as will be explained in Subsection 5.3.1. Further, the smoothing parameter is updated to obtain a new point. That is, if $\left\| \partial g \left(\mathbf{z}^{(t+1)}, \mu_{(t)} \right) \right\|_2 \geq \gamma \mu_{(t)}$, in Line 5 is not satisfied, then the smoothing parameter is updated using the new point in Line 8. Each vector $\partial g(\mathbf{z}^{(t)}, \mu_{(t)})$ in Algorithm 5 is calculated using the Wirtinger derivative as was introduced in Hunger (2007). SPRSF applies gradient

Algorithm 5 Sparse Phase Retrieval Algorithm via Smoothing Function (SPRSF)

1: **Input:** Data $\{(\mathbf{a}_i; q_i)\}_{i=1}^m$, sparsity level k. The step size $\tau \in (0,1)$, control variables $\gamma, \gamma_1 \in (0,1)$, $\mu_{(0)} \in \mathbb{R}_{++}$ and number of iterations T.
2:
3: **Initialization:** S_0 set to be the set of k largest indices of $\{\frac{1}{m}\sum_{i=1}^m q_i^2 a_{i,j}^2\}_{1\le j\le n}$. Let $\tilde{\mathbf{x}}^{(0)}$ be the leading eigenvector of the matrix $\mathbf{Y} := \frac{1}{m}\sum_{i\in I_0}\sqrt{q_i}\frac{\mathbf{a}_{i,S_0}\mathbf{a}_{i,S_0}^H}{\|\mathbf{a}_{i,S_0}\|_2^2}$. Define the initial point as $\mathbf{z}^{(0)} := \lambda_0\tilde{\mathbf{x}}^{(0)}$, where $\lambda_0 := \sqrt{\frac{\sum_{i=1}^m q_i^2}{m}}$.
4:
5: **for** $t = 0 : T-1$ **do**
6: $\quad \mathbf{z}^{(t+1)} = \mathscr{H}_k(\mathbf{z}^{(t)} - \tau\partial g(\mathbf{z}^{(t)}, \mu_{(t)}))$
7: $\quad$ **if** $\|\partial g\left(\mathbf{z}^{(t+1)}, \mu_{(t)}\right)\|_2 \ge \gamma\mu_{(t)}$ **then**
8: $\quad\quad \mu_{(t+1)} = \mu_{(t)}$
9: $\quad$ **else**
10: $\quad\quad \mu_{(t+1)} = \gamma_1\mu_{(t)}$
11: $\quad$ **end if**
12: **end for**
13: **Output:** $\mathbf{z}^{(T)}$

iterations based on the Wirtinger derivative to refine the initial estimate. Specifically, the Wirtinger derivative of $g(\mathbf{z}^{(t)}, \mu_{(t)})$ is given by

$$\partial g(\mathbf{z}^{(t)}, \mu_{(t)}) = \frac{2}{m}\sum_{i=1}^m \left(\mathbf{a}_i^H\mathbf{z}^{(t)} - q_i\frac{\mathbf{a}_i^H\mathbf{z}^{(t)}}{\sqrt{|\mathbf{a}_i^H\mathbf{z}^{(t)}|^2 + \mu_{(t)}^2}}\right)\mathbf{a}_i. \tag{88}$$

Notice that, in contrast to the gradient update steps for the SPARTA method introduced in Wang et al. (2016b), $\partial g(\mathbf{z}^{(t)}, \mu_{(t)})$ in (88) is always continuous because $\mu_{(t)} \neq 0$ for any $t \in \mathbb{N}$. Therefore, the proposed SPRSF method does not require any truncation parameter.

5.3.1. *Thresholded Gradient Stage*. The proposed Algorithm 5 solves the sparsity constraint of the optimization problem in (65) by iteratively refining the current update step $\mathbf{z}^{(t)}$ by a k-sparse hard thresholding operator $\mathscr{H}_k(\cdot)$, as calculated in Line 4 in Algorithm 5. Specifically,

$\mathscr{H}_k(\mathbf{u})$ sets all the entries in the vector $\mathbf{u} \in \mathbb{C}^n$ to zero, except for its k largest absolute values.

5.3.2. *Convergence Conditions*. This subsection provides theoretical results that guarantee the convergence of the proposed method summarized in Algorithm 5. The following theorem establishes that the successive estimates of SPRSF in Line 4 of Algorithm 5, tend to the unknown desired signal $\mathbf{x} \in \mathbb{C}^n$ for a given value of μ.

Theorem 5. (*Local error contraction*): Let $\mathbf{x} \in \mathbb{C}^n$ be any k-sparse ($k \ll n$) signal vector with the minimum nonzero entry on $(1/\sqrt{k})\|\mathbf{x}\|_2$. Consider the measurements $q_i = |\langle \mathbf{a}_i, \mathbf{x} \rangle|$, where $\mathbf{a}_i \sim \mathscr{CN}(0, \mathbf{I}_n), \forall i = 1, \cdots, m$. With a constant step size $\tau \in (0,1)$, successive estimates of SPRSF in Algorithm 3 satisfy

$$d_r(\mathbf{z}^{(t+1)}, \mathbf{x}) \leq \delta(1-\eta)^{t+1}\|\mathbf{x}\|_2 \tag{89}$$

which holds with probability exceeding $1 - 2e^{-c_1 m}$ provided that $m \geq C_1 k^2 \log(mn)$. Here, $c_1, C_1 \geq 0$ and $0 < \eta < 1$ are some universal constants. The constant δ is obtained from (17)

Demostración. The proof of Theorem 5 can be found in Appendix C. □

Note that Theorem 5 provides that the sequence $\{\mathbf{z}^{(t)}\}_{t\geq 1}$, generated by Algorithm 3, produces a monotonically decreasing sequence $\{g(\mathbf{z}^{(t)}, \mu)\}_{t\geq 1}$, with a given μ. Moreover, the sampling complexity bound $m \geq C_1 k^2 \log(mn)$, can often be rewritten as $m \geq C_1' k^2 \log(n)$ for some constant $C_1' \geq C_1$ and large enough n Wang et al. (2016b). Thus, it can be concluded that the sampling complexity of the SPRSF algorithm is $\mathscr{O}(k^2 \log(n))$.

5.4. Smoothing Phase Retrieval with Outliers

In some applications, it is expected to have measurements corrupted by sparse outliers Shuyue and Hongnian (2000). These arise due to various factors such as illumination, occlusion, device malfunctioning, or simply recording errors Chen et al. (2017). For instance, in X-ray imaging, the noise produced by the intensity of the X-ray radiation or charge-couple devices (CCD) exposure time is not large enough to be treated as Gaussian noise and should be modeled as outliers Qian et al. (2017); Shuyue and Hongnian (2000). Nevertheless, under this scenario, the methods mentioned above for PR have not shown good performance. Therefore, the PR problem needs to change its formulation. The phase retrieval problem contaminated by sparse arbitrary outliers can be formulated as a system of m quadratic equations of the form

$$y_k = |\mathbf{a}_k^H \mathbf{x}|^2 + o_k, \quad k = 1, \cdots, m, \tag{90}$$

where $o = [o_1, \cdots, o_m]$ is a sparse vector with αm non-zero entries and $(\cdot)^H$ denotes the conjugate transpose operation. To solve the phase retrieval with outliers, this thesis is based on the proposed smoothing method where first is necessary a proper initialization and then refined by gradient descent steps.

Considering that (61) is a smooth optimization problem for $\mu > 0$, now a gradient descent

procedure to solve (65) can be applied. To remember, the gradient of $g(\cdot)$ is defined by

$$\partial g(\mathbf{x},\mu) = \frac{2}{m}\sum_{k=1}^{m}\left(\frac{\varphi_\mu\left(|\mathbf{a}_k^H\mathbf{x}|\right)-\sqrt{y_k}}{\varphi_\mu\left(|\mathbf{a}_k^H\mathbf{x}|\right)}\right)\mathbf{a}_k\mathbf{a}_k^H\mathbf{x}. \tag{91}$$

Note that the gradient depends on the term $\varphi_\mu\left(|\mathbf{a}_k^H\mathbf{x}|\right)-\sqrt{y_k}$, which has a direct relationship with outliers expressed as

$$c|o_{k_t}| = |\varphi_\mu\left(|\mathbf{a}_{k_t}^H\mathbf{x}|\right)-\sqrt{y_{k_t}}| \tag{92}$$

for a constant c and k_t in the support of $\mathbf{o}$. In a gradient descent method, these positions drastically modify the search direction update. Therefore, in order to remove the corrupted components of the gradient in (91), an adaptive truncated version of $\partial g(\mathbf{x}_{(t)},\mu_{(t)})$ is calculated in each iteration. Therefore, the proposed procedure effectively dealing with outliers is given by

$$\mathbf{z}_{(t+1)} = \mathbf{z}_{(t)} - \frac{2\lambda}{m}\sum_{k\in\mathscr{T}_{(t)}}\partial g(\mathbf{z}_{(t)},\mu_{(t)}), \tag{93}$$

where $\mathscr{T}_{(t)}$ are given by the residual using the current iteration as

$$\begin{aligned}\mathscr{T}_{(t)} :=& \{k : |\left|\varphi_\mu\left(|\mathbf{a}_k^H\mathbf{z}_{(t)}|\right)-\sqrt{y_k}\right| \\ &\leq \beta\ \mathrm{med}(\{\left|\varphi_\mu\left(|\mathbf{a}_k^H\mathbf{z}_{(t)}|\right)-\sqrt{y_k}\right|\}_{k=1}^m)\}\end{aligned} \tag{94}$$

where β is a regularization parameter and $\mathrm{med}(\cdot)$ denotes the sample mean operator. Notice

that (94) prune samples whose residual gradient components are much large than the sample median. This robust property of median lies in the fact that the median cannot be affected with large values of outliers. Algorithm 6 summarizes the proposed method, where the initial estimation is calculated in line 2; lines 3 and 4 refine this initial approximation and finally, the μ parameter is updated in lines 6-10.

Notice that using the median as an estimator, $\mathcal{O}(n \log n)$ measurements are needed. Specifically, Zhang et al. (2018) provides a theoretical result that shows that it achieves recovery based on the local error contraction, which ensures that

$$dist(\mathbf{z}_{(t)}, \mathbf{x}) = \upsilon(1-\rho)^{(t)}||\mathbf{x}||_2, \tag{95}$$

for $0 < \rho,\ \upsilon < 1$ with probability at least $1 - c_1 exp(c_2 m)$ for $c_1, c_2 > 0$ if $m > n \log n$ measure-

Algorithm 6 RSPR: Robust smoothing phase retrieval algorithm

1: **Input:** Data $\{(\mathbf{a}_k; y_k)\}_{k=1}^m$ and choose constants $\beta = 4{,}6$. Choose $\gamma_1 \in (0,1), \mu_0 > 0, \gamma > 0$ and T maximum number of iterations.
2: Initial point $\mathbf{x}_{(0)} = \lambda_0 \tilde{\mathbf{z}}_0$, where $\lambda_0 = \sqrt{\frac{n\sum_{k=1}^m y_k}{\sum_{k=1}^m ||\mathbf{a}_k||^2}}$ and $\tilde{\mathbf{z}}_0$ is the leading eigenvector of $\mathbf{Y}_0 := \frac{1}{|I_0|}\sum_{k\in I_0} \frac{\mathbf{a}_k \mathbf{a}_k^H}{\|\mathbf{a}_k\|_2^2}$ where I_0 contains the indices corresponding to the $|I_0| = \frac{m}{6}$ largest values of $\{y_k/\|\mathbf{a}_k\|_2\}$.
3: **for** $t = 0 : T-1$ **do**
4: $\mathscr{T}_{(t)} := \{k : \left|\left|\varphi_\mu\left(|\mathbf{a}_k^H \mathbf{z}_{(t)}|\right) - \sqrt{y_k}\right| \leq \beta \text{ med}(\{\left|\varphi_\mu\left(|\mathbf{a}_k^H \mathbf{z}_{(t)}|\right) - \sqrt{y_k}\right)\}_{k=1}^m)\}$
5: $\mathbf{z}_{(t+1)} = \mathbf{z}_{(t)} - \frac{2\lambda}{m}\sum_{k\in\mathscr{T}_{(t)}} \left(1 - \frac{\sqrt{y_k}}{\varphi_\mu\left(|\mathbf{a}_k^H \mathbf{z}_{(t)}|\right)}\right) \mathbf{a}_k \mathbf{a}_k^H \mathbf{z}_{(t)}$
6: **if** $||\partial g(\mathbf{z}_{(t+1)}, \mu_{(t)})||_2 \leq \gamma \mu_{(t)}$ **then**
7: $\mu_{(t+1)} = \gamma_1 \mu_{(t)}$
8: **else**
9: $\mu_{(t+1)} = \mu_{(t)}$
10: **end if**
11: **end for**
12: **return:** $\mathbf{z}_{(T)}$

ments are available. Also, because of the geometric convergence rate, the proposed method achieves ε-accuracy (i.e. $dist(\mathbf{z}_{(t)}, \mathbf{x}) = \varepsilon||x||_2$) within at most $\mathscr{O}(\log(1/\varepsilon))$ iterations. Therefore, the computation cost is $\mathscr{O}(mn \log(1/\varepsilon))$.

5.5. Super Resolution Phase Retrieval Algorithm

This section presents a reconstruction algorithm focused on the super resolution scenario proposed in this thesis to estimate a high-resolution image from the low-resolution measurements $\mathbf{g}_u$ acquired at the u-th diffraction zone. However, as shown below, the split and sparsity methodology can also be used for super-resolution scenarios and the traditional phase retrieval problem. This method follows the smoothing technique introduced in Pinilla et al. (2018a) to overcome the non-smoothness of the amplitude-based objective. Specifically, using Euler's formula, $\mathbf{x} = \mathbf{r} \odot e^{j\varphi}$, where $\mathbf{r}$, φ are the magnitude and phase of $\mathbf{x}$ and $\odot$ denotes the Hadamard product, the smooth objective takes the form

$$\begin{aligned} \min_{\mathbf{r},\varphi,\mathbf{x}\in\mathbb{C}^n} \quad & \frac{1}{Lm}\sum_{i=1}^{mL}\left(\vartheta_\mu(|\mathbf{a}_{u,i}^H\mathbf{x}|) - \sqrt{(\mathbf{g}_u)_i}\right)^2 + \lambda_{tv_1}\|\mathbf{L}\mathbf{r}\|_1 \\ & + \lambda_{tv_2}\|\mathbf{L}\varphi\|_1 \end{aligned}$$

$$\text{subject to} \quad \mathbf{x} - \mathbf{r} \odot e^{j\varphi} = \mathbf{0}, \tag{96}$$

where the function $\vartheta_\mu : \mathbb{R}_+ \to \mathbb{R}_+$, defined as $\vartheta_\mu(w) = \sqrt{w^2+\mu^2}$ with $\mu \in \mathbb{R}_{++}$, smooths the term $|\mathbf{a}_{u,i}^H \mathbf{x}|$.

Notice that in (96) the spatial property of the magnitude and phase of $\mathbf{x}$ is exploited by minimizing the total variation (TV) where λ_{tv_1} and λ_{tv_2} are the regularization parameters associated with TV, and $\mathbf{L} = [\mathbf{L}_h; \mathbf{L}_v]$ denotes the operator that computes the horizontal and vertical differences. Specifically, these two regularization terms come from the fact that both, $\mathbf{r}$ and φ should be spatially continuous Katkovnik and Egiazarian (2017).

It is important to note that (96) is non-convex with respect to $\mathbf{x}$, $\mathbf{r}$, and φ, making this problem more challenging. However, (96) has solution with respect to $\mathbf{x}$, $\mathbf{r}$, and φ separately. Therefore, this thesis proposes an algorithm that starts with a proper initialization, and the optimization problem in (96) is solved one vector at a time, while the other variables are assumed to be fixed, as summarized in Algorithm 7. More details about every step of Algorithm 7 are provided below.

5.5.1. ***Optimization with respect to*** **x.** Considering the optimization problem in (96), the minimization problem with respect to $\mathbf{x}$ can be expressed as

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{C}^n} \; f(\mathbf{x},\mu) = \frac{1}{Lm}\sum_{i=1}^{mL}\left(\vartheta_\mu(|\mathbf{a}_{u,i}^H\mathbf{x}|) - \sqrt{(\mathbf{g}_u)_i}\right)^2 \\ + \frac{\rho}{2}\|\mathbf{x} - \mathbf{r}^{(t)} \odot e^{j\varphi^{(t)}}\|_2^2, \end{aligned} \tag{97}$$

where $\rho > 0$ is a regularization parameter. This sub-step can be solved with a descent gradient method based on the Wirtinger derivative as introduced in Candes et al. (2015d), which in this case

Algorithm 7 Super-Resolution Phase Retrieval Algorithm

Input: Data $\{(\mathbf{a}_{u,i};(\mathbf{g}_u)_i)\}_{i=1}^{mL}$, $\mu_0 \in \mathbb{R}_{++}$, and number of iterations T.
Initial point $\mathbf{x}^{(0)} = \sqrt{\frac{\sum_{i=1}^{mL}(\mathbf{g}_u)_i}{mL}}\tilde{\mathbf{x}}^{(0)}$, where $\tilde{\mathbf{x}}^{(0)}$ is the leading eigenvector of $\mathbf{Y}_0 := \frac{1}{|I_0|}\sum_{i\in I_0}\frac{\mathbf{a}_{u,i}\mathbf{a}_{u,i}^H}{\|\mathbf{a}_{u,i}\|_2^2}$ given by the power iteration method.
$\mathbf{r}^{(0)} = \left|\mathbf{x}^{(0)}\right|$
$\varphi^{(0)} = \text{Phase}(\mathbf{x}^{(0)})$.
for $t = 0 : T-1$ **do**
  $\mathbf{x}^{(t+1)} = \text{Algorithm 2 } (\mathbf{r}^{(t)}, \varphi^{(t)})$
  $\mathbf{r}_{\mathbf{x}^{(t+1)}} = \left|\mathbf{x}^{(t+1)}\right|$
  $\varphi_{\mathbf{x}^{(t+1)}} = \text{Phase}(\mathbf{x}^{(t+1)})$
  $\mathbf{r}^{(t+1)} = \text{Algorithm 3 } (\mathbf{r}_{\mathbf{x}^{(t+1)}})$
  $\varphi^{(t+1)} = \text{Algorithm 4 } (\varphi_{\mathbf{x}^{(t+1)}})$
end for
$\mathbf{w} = \mathbf{r}^{(T)} \odot e^{j\varphi^{(T)}}$
Output: **w**

is given by

$$\partial f(\mathbf{x},\mu) = \frac{1}{Lm}\sum_{i=1}^{mL}\left(\mathbf{a}_{u,i}^H\mathbf{x} - \sqrt{(\mathbf{g}_u)_i}\frac{\mathbf{a}_{u,i}^H\mathbf{x}}{\vartheta_\mu(|\mathbf{a}_{u,i}^H\mathbf{x}|)}\right)\mathbf{a}_{u,i} + \rho(\mathbf{x} - \mathbf{r}^{(t)} \odot e^{j\varphi^{(t)}}). \tag{98}$$

Additional, in each iteration the smoothing parameter is updated to obtain a new point if $\|\partial f\left(\mathbf{x}^{(s+1)},\mu^{(s)}\right)\|_2 \leq \gamma\mu^{(s)}$ is satisfied. The solution of this problem is summarized in Algorithm 8.

Convergence and computational complexity: Since Algorithm 8 is a gradient descent method, $\partial f(\mathbf{x},\mu)$ needs to be Lipschitz continuous Grippo et al. (1986). This condition is satisfied if the Wirtinger derivative of the two terms in (97) are Lipschitz continuous. Specifically, Theorem

Algorithm 8 Algorithm to estimate $\mathbf{x}$

Input: Previous iterations $\mathbf{r}^{(t)}$, and $\varphi^{(t)}$.
Initialize: Constants $\tau, \gamma, \gamma_1 \in (0,1)$, and the number of iterations S_1.
for $s = 0 : S_1 - 1$ **do**
  $\mathbf{x}^{(s+1)} \leftarrow \mathbf{x}^{(s)} - \tau \partial f(\mathbf{x}^{(s)}, \mu^{(s)})$
  if $\|\partial f\left(\mathbf{x}^{(s+1)}, \mu^{(s)}\right)\|_2 \leq \gamma \mu^{(s)}$ **then**
    $\mu^{(s+1)} = \gamma_1 \mu^{(s)}$
  else
    $\mu^{(s+1)} = \mu^{(s)}$
  end if
end for
Output: $\mathbf{x}^{(S_1)}$

2 from Pinilla et al. (2018a) proved that the Wirtinger derivative of the first term of $f(\mathbf{x}, \mu)$ is Lipschitz continuous. Also, given the fact that the second term in (97) is based on the ℓ_2 norm, this condition is trivially satisfied. Thus, the convergence of Algorithm 8 is guaranteed. Finally, following the iteration process of Algorithm 8, it can be seen that its computational complexity is $\mathcal{O}(LS_1)$.

5.5.2. *Optimization with respect to* r. To obtain the update step of $\mathbf{r}$, the cost function in (96) is minimized with respect to $\mathbf{r}$ resulting in the following optimization problem

$$\begin{aligned} \min_{\mathbf{r} \in \mathbb{R}^n} \quad & \lambda_{tv_1} \|\mathbf{L}\mathbf{r}\|_1 \\ \text{subject to} \quad & \mathbf{r}_{\mathbf{x}^{(t+1)}} - \mathbf{r} = \mathbf{0}, \end{aligned} \tag{99}$$

where $\mathbf{r}_{\mathbf{x}^{(t+1)}}$ is the magnitude of $\mathbf{x}^{(t+1)}$ at the global iteration $t+1$. In order to solve (99), an alternating direction method of multipliers (ADMM) strategy is used. Specifically, an auxiliary

variable $\mathbf{z}$ is introduced, such that the optimization problem in (99) can be rewritten as

$$\begin{aligned} \underset{\mathbf{r},\mathbf{z}\in\mathbb{R}^n}{\text{mín}} \quad & \lambda_{tv_1}\|\mathbf{z}\|_1 \\ \text{subject to} \quad & \mathbf{r}_{\mathbf{x}^{(t+1)}} - \mathbf{r} = \mathbf{0} \\ & \mathbf{z} - \mathbf{L}\mathbf{r} = \mathbf{0}. \end{aligned} \tag{100}$$

Note that the augmented Lagrangian associated to the optimization problem in (100) is expressed as

$$\begin{aligned} \mathscr{L}(\mathbf{r},\mathbf{z},\mathbf{g}_1,\mathbf{g}_2) = \lambda_{tv_1}\|\mathbf{z}\|_1 + \frac{\rho}{2}\|\mathbf{r}_{\mathbf{x}^{(t+1)}} - \mathbf{r} + \mathbf{g}_1\|_2^2 \\ + \frac{\rho}{2}\|\mathbf{z} - \mathbf{L}\mathbf{r} + \mathbf{g}_2\|_2^2, \end{aligned} \tag{101}$$

where $\mathbf{g}_1$ and $\mathbf{g}_2$ are the scaled dual variables, and $\rho > 0$ is the weighting of the augmented Lagrangian term. The solution for each variable of $\mathscr{L}(\cdot)$ is summarized in Algorithm 9. Specifically,

Algorithm 9 Algorithm to estimate $\mathbf{r}$

Input: vector $\mathbf{r}_{\mathbf{x}^{(t)}}$
Initialize: constants $\rho, \lambda_{tv_1} > 0$, and the number of iterations S_2
$\mathbf{z} = \mathbf{g}_1 = \mathbf{g}_2 = \mathbf{0}$
for $s = 0 : S_2 - 1$ **do**
  $\mathbf{r}^{(s+1)} = \left(\mathbf{L}^T\mathbf{L} + \mathbf{I}\right)^{-1}\left(\mathbf{r}_{\mathbf{x}^{(t+1)}} + \mathbf{g}_1^{(s)} + \mathbf{L}^T(\mathbf{z}^{(s)} + \mathbf{g}_2^{(s)})\right)$
  $\mathbf{z}^{(s+1)} = \mathscr{S}_{\lambda_{tv_1}/\rho}\left(\mathbf{L}\mathbf{r}^{(s+1)} - \mathbf{g}_2^{(s)}\right)$
  $\mathbf{g}_1^{(s+1)} = \mathbf{g}_1^{(s)} + \mathbf{r}_{\mathbf{x}^{(t+1)}} - \mathbf{r}^{(s+1)}$
  $\mathbf{g}_2^{(s+1)} = \mathbf{g}_2^{(s)} + \mathbf{z}^{(s+1)} - \mathbf{L}\mathbf{r}^{(s+1)}$
end for
Output: $\mathbf{r}^{(S_2)}$

Algorithm 9 begins by initializing the variables $\mathbf{z} = \mathbf{g}_1 = \mathbf{g}_2 = \mathbf{0}$. Then the Lagrangian with respect to the variable $\mathbf{r}$ is minimized. The solution for $\mathbf{r}$ is given by

$$\begin{aligned} \mathbf{r}^{(s+1)} &= \underset{\mathbf{r}\in\mathbb{R}^n}{\arg\min} \;\; \|\mathbf{r}_{\mathbf{x}^{(t+1)}} - \mathbf{r} + \mathbf{g}_1^{(s)}\|_2^2 + \|\mathbf{z}^{(s)} - \mathbf{L}\mathbf{r} + \mathbf{g}_2^{(s)}\|_2^2 \\ &= \left(\mathbf{L}^T\mathbf{L} + \mathbf{I}\right)^{-1}\left(\mathbf{r}_{\mathbf{x}^{(t+1)}} + \mathbf{g}_1^{(s)} + \mathbf{L}^T(\mathbf{z}^{(s)} + \mathbf{g}_2^{(s)})\right), \end{aligned} \tag{102}$$

as computed in Line 5. On the other hand, the solution for $\mathbf{z}$ in Line 6 is computed as

$$\begin{aligned} \mathbf{z}^{(s+1)} &= \underset{\mathbf{z}\in\mathbb{R}^n}{\arg\min} \;\; \lambda_{tv_1}\|\mathbf{z}\|_1 + \frac{\rho}{2}\|\mathbf{z} - \mathbf{L}\mathbf{r}^{(s+1)} + \mathbf{g}_2^{(s)}\|_2^2 \\ &= \mathscr{S}_{\lambda_{tv_1}/\rho}\left(\mathbf{L}\mathbf{r}^{(s+1)} - \mathbf{g}_2^{(s)}\right), \end{aligned} \tag{103}$$

where $\mathscr{S}_\tau(\cdot)$ denotes the component-wise application of the soft-threshold function as

$$\mathscr{S}_\tau(\mathbf{w}) = \text{sign} \odot \text{máx}(0, |\mathbf{w}| - \tau), \tag{104}$$

Finally the updates of the dual variables for the iteration $s+1$ are given by

$$\mathbf{g}_1^{(s+1)} = \mathbf{g}_1^{(s)} + \mathbf{r}_{\mathbf{x}^{(t+1)}} - \mathbf{r}^{(s+1)} \tag{105}$$

$$\mathbf{g}_2^{(s+1)} = \mathbf{g}_2^{(s)} + \mathbf{z}^{(s+1)} - \mathbf{L}\mathbf{r}^{(s+1)}, \tag{106}$$

as summarized in Lines 7 and 8 of Algorithm 9.

5.5.3. *Optimization with respect to* φ. Finally, the update step of φ, considering (112), can be obtained solving the following optimization problem

$$\begin{aligned} \min_{\varphi\in\mathbb{R}^n} \quad & \lambda_{tv_2}\|\mathbf{L}\varphi\|_1 \\ \text{subject to} \quad & \varphi_{\mathbf{x}^{(t+1)}} - \varphi = \mathbf{0}, \end{aligned} \tag{107}$$

where $\varphi_{\mathbf{x}^{(t+1)}}$ is the phase of $\mathbf{x}^{(t+1)}$ at the global iteration $t+1$. To solve (107), an ADMM strategy was followed and summarized in Algorithm 10. Also, notice that (99) and (107) are similar optimization problems. Performing a similar derivation for φ, from (102) the solution for φ is given by

$$\begin{aligned} \varphi^{(s+1)} &= \arg\min_{\varphi\in\mathbb{R}^n} \ \|\varphi_{\mathbf{x}^{(t+1)}} - \varphi + \mathbf{d}_1^{(s)}\|_2^2 + \|\mathbf{b}^{(s)} - \mathbf{L}\varphi + \mathbf{d}_2^{(s)}\|_2^2 \\ &= \left(\mathbf{L}^T\mathbf{L} + \mathbf{I}\right)^{-1}\left(\varphi_{\mathbf{x}^{(t+1)}} + \mathbf{d}_1^{(s)} + \mathbf{L}^T(\mathbf{b}^{(s)} + \mathbf{d}_2^{(s)})\right), \end{aligned} \tag{108}$$

as computed in Line 5 of Algorithm 10, where $\mathbf{d}_1$ and $\mathbf{d}_2$ are the scaled dual variables, and $\mathbf{b}$ is an auxiliary variable. On the other hand, the solution for $\mathbf{b}$ is similarly written as

$$\begin{aligned} \mathbf{b}^{(s+1)} &= \arg\min_{\mathbf{b}\in\mathbb{R}^n} \ \lambda_{tv_2}\|\mathbf{b}\|_1 + \frac{\sigma}{2}\|\mathbf{b} - \mathbf{L}\varphi^{(s+1)} + \mathbf{d}_2^{(s)}\|_2^2 \\ &= \mathcal{S}_{\lambda_{tv_2}/\sigma}\left(\mathbf{L}\varphi^{(s+1)} - \mathbf{d}_2^{(s)}\right), \end{aligned} \tag{109}$$

as calculated in Line 6 of Algorithm 10, where $\sigma > 0$ is the weighting of the augmented Lagrangian term. The final two steps of Algorithm 10 update the dual variables as follows

$$\mathbf{d}_1^{(s+1)} = \mathbf{d}_1^{(s)} + \varphi_{\mathbf{x}^{(t+1)}} - \varphi^{(s+1)} \tag{110}$$

$$\mathbf{d}_2^{(s+1)} = \mathbf{d}_2^{(s)} + \mathbf{b}^{(s+1)} - \mathbf{L}\varphi^{(s+1)}, \tag{111}$$

as summarized in Lines 7 and 8.

Algorithm 10 Algorithm to estimate φ

1: **Input:** Vector $\varphi_{\mathbf{x}^{(t+1)}}$.
2: **Initialize:** Constants $\sigma, \lambda_{tv_2} > 0$, and the number of iterations S_2.
3: $\mathbf{b} = \mathbf{d}_1 = \mathbf{d}_2 = \mathbf{0}$.
4: **for** $s = 0 : S_2 - 1$ **do**
5: $\varphi^{(s+1)} = \left(\mathbf{L}^T\mathbf{L} + \mathbf{I}\right)^{-1} \left(\varphi_{\mathbf{x}^{(t+1)}} + \mathbf{d}_1^{(s)} + \mathbf{L}^T(\mathbf{b}^{(s)} + \mathbf{d}_2^{(s)})\right)$
6: $\mathbf{b}^{(s+1)} = \mathscr{S}_{\lambda_{tv_2}/\sigma}\left(\mathbf{L}\varphi^{(s+1)} - \mathbf{d}_2^{(s)}\right)$
7: $\mathbf{d}_1^{(s+1)} = \mathbf{d}_1^{(s)} + \varphi_{\mathbf{x}^{(t+1)}} - \varphi^{(s+1)}$
8: $\mathbf{d}_2^{(s+1)} = \mathbf{d}_2^{(s)} + \mathbf{b}^{(s+1)} - \mathbf{L}\varphi^{(s+1)}$
9: **end for**
10: **Output:** $\varphi^{(S_2)}$

Convergence and computational complexity: In order to guarantee the convergence of Algorithms 9 and 10, the augmented Lagrangian in (101) needs to be a proper convex and closed function, according to the ADMM algorithms. This condition is satisfied since (101) is the sum of non-negative convex functions Boyd et al. (2011). Moreover, since the proper convex optimization function is continuous, it is closed. Thus, the convergence of Algorithms 9 and 10 is guaranteed Boyd and Vandenberghe (2004). Following the iteration process of Algorithms 3, 4, and assuming that $(\mathbf{L}^T\mathbf{L} + \mathbf{I})^{-1}$ can be precomputed, it can be seen that the computational complexity of both

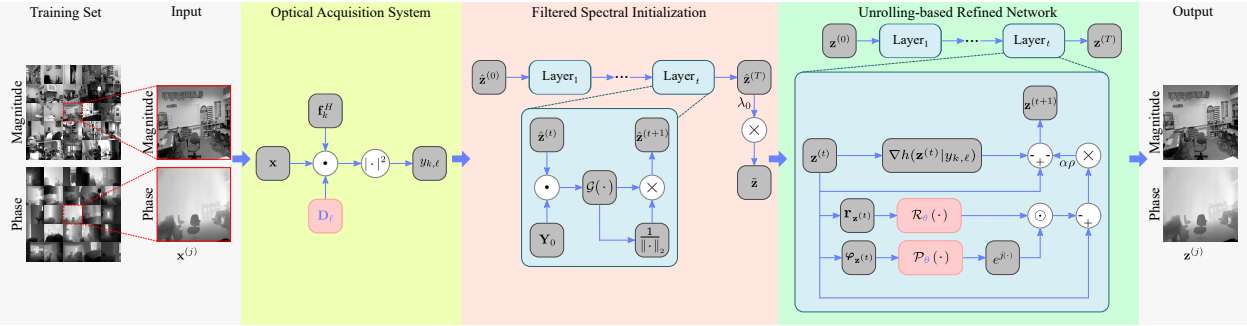
algorithms is $\mathscr{O}(S_2 n^2)$.

5.5.4. *Global Convergence*. Given that the optimization problem in (112), viewed as a function of $\mathbf{x}$, $\mathbf{r}$ or φ, attains a stationary point from Theorem 4.1 in Tseng (2001), every limit point of the sequence $\{\mathbf{x}^{(t)}, \mathbf{r}^{(t)}, \varphi^{(t)}\}$ generated by Algorithm 7 is a stationary point of the considered optimization problem. Finally, considering the computational complexity of Algorithms 8, 9, and 10, it can be concluded that the overall reconstruction process summarized in Algorithm 7 has a computational complexity $\mathscr{O}(LS_1 + S_2 n^2)$, which directly depends on the image size, number of experiments, and the iterations.

5.6. Deep Unrolled Recovery Network

Hoover, in some applications, the prior knowledge of the scene is not easy to incorporate as a regularizer; therefore, it is more convenient to learn the prior using the available data. Therefore, we designed a proposed unrolled decoder is inspired by the optimization formulation present in Bacca et al. (2018) which introduced a non-smooth function and exploited sparsity assumption for the magnitude and phase separately. Specifically, using Euler's formula, $\mathbf{x} = \mathbf{r} \odot e^{j\varphi}$, where $\mathbf{r}$, φ

Figure 9. Propposed E2E approach.



are the magnitude and phase of $\mathbf{x}$ and $\odot$ denotes the Hadamard product, the refined network aims to solve the following optimization problem

$$\begin{aligned} \min_{\mathbf{r},\varphi,\mathbf{z}\in\mathbb{C}^n} \quad & \frac{1}{2nL}\sum_{\ell=1}^{L}\sum_{k=0}^{n-1}\left(\vartheta_\mu(|\mathbf{f}_k^H\mathbf{D}_\ell\mathbf{z}|)-\sqrt{y_{k,\ell}}\right)^2 \\ & \qquad +\lambda_{\mathbf{r}}R_{\mathbf{r}}(\mathbf{r})+\lambda_\varphi R_\varphi(\varphi) \\ \text{subject to} \quad & \mathbf{z}-\mathbf{r}\odot e^{j\varphi}=\mathbf{0}, \end{aligned} \tag{112}$$

where the function $\vartheta_\mu : \mathbb{R}_+ \to \mathbb{R}_+$, defined as $\vartheta_\mu(w) = \sqrt{w^2+\mu^2}$ and $R_{\mathbf{r}}(\cdot)$ and $R_\varphi(\cdot)$ denoted the regularization of the magnitude and phase with its corresponding regularization terms $\lambda_{\mathbf{r}}$ and λ_φ, respectively. This thesis proposed to replace the effect of the regularization terms by a prior network as explained below. Particularly, the minimization respect of $\mathbf{z},\mathbf{r}$ and φ can be found by iteratively solving the following optimization problems

$$\begin{aligned} \mathbf{z}^{(t+1)} := \arg\min_{\mathbf{z}\in\mathbb{C}^n} & \frac{1}{2nL}\sum_{\ell=1}^{L}\sum_{k=0}^{n-1}\left(\vartheta_\mu(|\mathbf{f}_k^H\mathbf{D}_\ell\mathbf{z}|)-\sqrt{y_{k,\ell}}\right)^2 \\ & +\frac{\rho}{2}\|\mathbf{z}-\mathbf{r}^{(t)}\odot e^{j\varphi^{(t)}}\|_2^2, \end{aligned} \tag{113}$$

$$r^{(t+1)} := \arg\min_{\mathbf{r}\in\mathbb{C}^n} \lambda_r R_r(\mathbf{r}) + \frac{\rho}{2}\|\mathbf{r}_{\mathbf{z}^{(t+1)}}-\mathbf{r}\|_2^2, \tag{114}$$

$$\varphi^{(t+1)} := \arg\min_{\varphi\in\mathbb{C}^n} \lambda_\varphi R_\varphi(\varphi) + \frac{\rho}{2}\|\varphi_{\mathbf{z}^{(t+1)}}-\varphi\|_2^2, \tag{115}$$

where $\mathbf{r}_{\mathbf{z}^{(t+1)}}$ and $\varphi_{\mathbf{z}^{(t+1)}}$ are the magnitude and phase of $\mathbf{z}^{(t+1)}$ at the global iteration $t+1$. this thesis solves (113) using a descent gradient method based on the Wirtinger derivative as

$$\begin{aligned} z^{(t+1)} := & \\ & z^{(t)} - \frac{\alpha}{Ln} \sum_{\ell=1}^{L} \sum_{k=0}^{n-1} \left(\mathbf{f}_k^H \mathbf{D}_\ell \mathbf{z}^{(t)} - \frac{\sqrt{y_{k,\ell}} \mathbf{f}_k^H \mathbf{D}_\ell \mathbf{z}^{(t)}}{\vartheta_\mu(|\mathbf{f}_k^H \mathbf{D}_\ell \mathbf{z}^{(t)}|)} \right) \mathbf{D}^H \mathbf{f}_k \\ & - \alpha\rho \left(\mathbf{z} - \mathbf{r}^{(t)} \odot e^{j\varphi^{(t)}} \right), \end{aligned} \tag{116}$$

where $\alpha > 0$ is the gradient descent step size and $\rho > 0$ is a regularization parameter. The minimizing of $\mathbf{r}^{(t)}$ and $\varphi^{(t)}$ can be seen as a proximal operator that can be addressed by applying a DNN at the magnitude and phase of $\mathbf{z}^{(t+1)}$ as

$$\mathbf{r}^{(t+1)} = \mathscr{R}_\theta \left(\mathbf{r}_{\mathbf{z}^{(t+1)}} \right), \tag{117}$$

$$\varphi^{(t+1)} = \mathscr{P}_\theta \left(\varphi_{\mathbf{z}^{(t+1)}} \right), \tag{118}$$

where $\mathscr{R}_\theta$ and $\mathscr{P}_\theta$ represent a DNN with trainable parameters θ for the magnitude and phase, respectively. Finally, combing (116), (117) and (118) in a hold step, the image can be estimated as

$$\begin{aligned}\mathbf{z}^{(t+1)} :=& \\ &\mathbf{z}^{(t)} - \underbrace{\frac{\alpha}{Ln}\sum_{\ell=1}^{L}\sum_{k=0}^{n-1}\left(\mathbf{f}_k^H\mathbf{D}_\ell\mathbf{z}^{(t)} - \frac{\sqrt{y_{k,\ell}}\mathbf{f}_k^H\mathbf{D}_\ell\mathbf{z}^{(t)}}{\vartheta_\mu(|\mathbf{f}_k^H\mathbf{D}_\ell\mathbf{z}^{(t)}|)}\right)\mathbf{D}^H\mathbf{f}_k)}_{\nabla h(\mathbf{z}^{(t)}|y_{k,\ell}} \\ &- \alpha\rho\left(\mathbf{z}^{(t)} - \mathscr{R}_\theta\left(\mathbf{r}_{\mathbf{z}^{(t)}}\right)\odot e^{j\mathscr{P}_\theta\left(\varphi_{\mathbf{z}^{(t)}}\right)}\right).\end{aligned} \tag{119}$$

To solve (119), and find the optimal θ parameter, this thesis proposes to unroll the recursion iteration via DNN from $t \in \{1,\cdots,T\}$ step iterations, as depicted in Fig. 9, which results in T layers of the proposed method denoted as $\mathscr{M}_\theta(\cdot)$. Additionally, The parameters λ and μ are also trainable in the unrolled DNN.

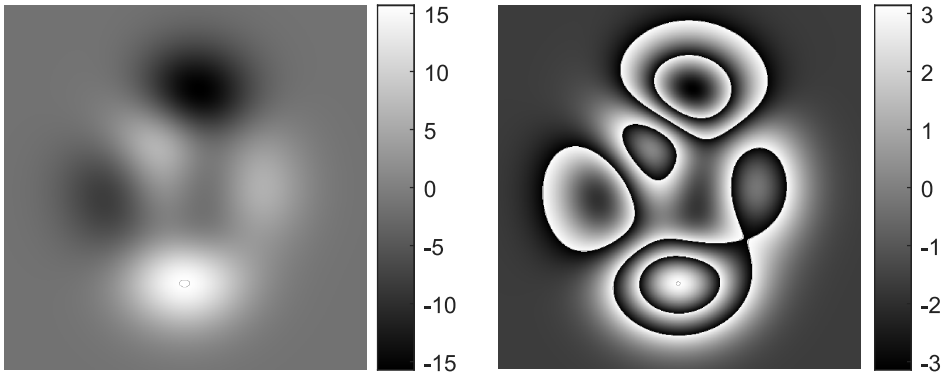
5.7. SPUD: simultaneous phase unwrapping and denoising algorithm for phase imaging

5.7.1. *Phase unwrapping*. When we work with continuous phase images, as is the focus of this thesis, it is necessary to apply unwrapping methods to use prior information to the phase image correctly. This is due to the natural periodicity of the phase image, where values between $[-\pi,\pi]$ or $[0,2\pi]$ only produce the different effect in the phase, i.e., a value of π make the same offset of a value 3π. For instance, Fig.10 presents a continuous phase image and its wrapped version, where both produce the same phase delay.

This is a common ambiguity present in the recovery algorithms. However, to explore the prior information of the image is convenient to work with the continuous phase image instead

of the wrapped version. For instance, it is easier to apply denoiser for the constant phase than the wrapped version Pineda et al. (2020). So, several phase retrieval algorithms employ unwrapping algorithms in each step of the algorithm before exploiting the prior information. However, unwrapping methods use iterative procedures to obtain the solution, resulting in long execution times Pineda et al. (2020). Therefore, this thesis presents a non-iterative Simultaneous Phase Unwrapping and Denoising algorithm for phase imaging, referred to as SPUD. The proposed method relies on the least-squares Discrete Cosine Transform (DCT) solution for phase unwrapping with an additional sparsity constraint on the DCT coefficients of the unwrapped solution.

Figure 10. (Left) Continuous phase image (Right) Wrapped phase image



Note: Both images produce the same phase delay.

5.7.2. *Problem Formulation*. The goal in 2D phase unwrapping is to estimate the true phase image $\phi \in \mathbb{R}^{M\times N}$, from a wrapped phase image $\varphi \in (-\pi,\pi]^{M\times N}$ defined by

$$\varphi = \mathscr{W}\{\phi\},' \tag{120}$$

where $\mathscr{W}\{\cdot\}$ is the wrapping operator that performs component-wise 2π modulo wrapping operation

$$\begin{aligned} \mathscr{W} =: \mathbb{R}^{M\times N} &\to (-\pi,\pi]^{M\times N} \quad , \\ \varphi &\to \operatorname{mod}(\phi+\pi, 2\pi)-\pi \quad . \end{aligned} \tag{121}$$

The proposed formulation relies on two assumptions. The first is that the desired unwrapped phase and the wrapped phase have the same local phase differences. Therefore, it is conventional to define

$$\Delta\varphi^{x}_{i,j} = \mathscr{W}\left\{\varphi_{i+1,j}-\varphi_{i,j}\right\}, \;\; \Delta\varphi^{y}_{i,j} = \mathscr{W}\left\{\varphi_{i,j+1}-\varphi_{i,j}\right\} \quad , \tag{122}$$

as the horizontal and vertical phase differences, respectively. This assumption has an exact solution by solving a least-squares algorithm (in the noiseless scenario) Ghiglia and Romero (1996); Ghiglia and Pritt (1998) or shows desirable results when the noise present in the differences does not exceed π, i.e., $|\varphi_{i+1,j}-\varphi_{i,j}+\eta_{i,j}|<\pi$, where $\eta_{i,j}$ is the noise of the horizontal differences. It occurs similarly for vertical differences. The second assumption of this thesis is that the true phase image is smooth Estrada et al. (2011). Therefore, it can be sparsified in a given transformation $\mathscr{T}(\cdot)$, i.e., $||\mathscr{T}(\phi)||_0 = k \ll MN$, where $||\mathbf{x}||_0 = |\{i : \mathbf{x}_i \neq 0\}|$ with $|\{\cdot\}|$ as the cardinality of a set, such that, the ℓ_0-norm counts the number of nonzero elements of $\mathbf{x}$. Additionally, in a noisy phase image, the sparsity property implies that the relevant information is concentrated in few coefficients, while the power of the noise remains white Yu and Sapiro (2011). Hence, a least-squares phase unwrapping

formulation incorporating these two assumptions can be expressed as

$$\arg\min_{\phi} \left\{ \sum_{ij} (\Delta\phi^x_{i,j} - \Delta\varphi^x_{i,j})^2 + \sum_{ij} (\Delta\phi^y_{i,j} - \Delta\varphi^y_{i,j})^2 \right\} + \|\mathscr{T}(\phi)\|_0 \,, \tag{123}$$

where $\Delta\phi^x_{i,j}$ and $\Delta\varphi^x_{i,j}$ denote the x-components of the unwrapped and wrapped phase gradients, respectively; $\Delta\phi^y_{i,j}$ and $\Delta\varphi^y_{i,j}$ are their y components counterparts.

5.7.3. ***Simultaneous Phase Unwrapping and Denoising Algorithm.*** This thesis proposes a non-iterative method to solve (123). In particular, notice that the left side of (123) is reduced to the Hunt's matrix formulation given by,

$$\left(\phi_{i+1,j} - 2\phi_{i,j} + \phi_{i-1,j}\right) + \left(\phi_{i,j+1} - 2\phi_{i,j} + \phi_{i,j-1}\right) = \rho_{i,j} \ , \tag{124}$$

where,

$$\rho_{i,j} = \left(\Delta\varphi^x_{i,j} - \Delta\varphi^x_{i-1,j}\right) + \left(\Delta\varphi^y_{i,j} - \Delta\varphi^y_{i,j-1}\right) \ . \tag{125}$$

Additionally, (125) can be interpreted as the discretization of Poisson's equation with Neumann boundary conditions

$$\nabla^2\phi_{i,j} = \rho_{i,j} \ , \tag{126}$$

where ∇^2 is the Laplacian operator. Therefore, applying the two-dimensional DCT on the $M \times N$ grid to both sides of (126) yields

$$\hat{\phi}_{i,j} = \frac{\hat{\rho}_{i,j}}{2[\cos(\pi i/M) + \cos(\pi j/N) - 2]}, \tag{127}$$

where $\hat{\phi}_{i,j} = \mathscr{T}(\phi_{i,j})$ and $\hat{\rho}_{i,j} = \mathscr{T}(\rho_{i,j})$ denote the 2-D forward DCT of $\phi_{i,j}$ and $\rho_{i,j}$, respectively. The sparsity information can be exploited using the element-wise hard-thresholding operator $\Theta^{\lambda}_{hard}(.)$, which can be directly applied in the sparse vector to reduce the noise Yu and Sapiro (2011), i.e, in the DCT domain defined as

$$\Theta^{\lambda}_{hard}(\hat{\phi}_{i,j}) = \begin{cases} 0 \text{ if } \mid \hat{\phi}_{i,j} \mid \leqslant \lambda \\ \\ \hat{\phi}_{i,j}, \text{ otherwise} \end{cases}. \tag{128}$$

Finally, the noise-free solution $\phi_{i,j}$ is obtained by the inverse DCT of (128), i.e., $\phi = \mathscr{T}^{-1}\left(\Theta^{\lambda}_{hard}(\hat{\phi}_{i,j})\right)$. Notice that the mean squared error (MSE) of the true phase and the threshold estimation ((128)), can be written as

$$\begin{aligned} &\mathbb{E}\left[\left\|\phi - \mathscr{T}^{-1}\left(\Theta^{\lambda}_{hard}(\hat{\phi}_{i,j})\right)\right\|_2^2\right] = \\ &\sum_{i,j:|\hat{\phi}_{i,j}|\leqslant\lambda} \left|\mathscr{T}(\phi_{i,j})\right|^2 + \sigma^2 \left|\{i,j : |\hat{\phi}_{i,j}| > \lambda\}\right|, \end{aligned} \tag{129}$$

where the first and second terms are the bias and variance of the threshold estimation, assuming Gaussian white noise with variance σ^2. Additionally, with a threshold parameter λ sufficiently close to $\sigma\sqrt{2\log(MN)}$, the MSE is comparable to that of an oracle projection, which reduces the

variance without increasing the bias, resulting in an optimal denoising value Donoho and Johnstone (1994). Algorithm 1 summarizes the steps explained above.

Algorithm 11 SPUD: Simultaneous Phase Unwrapping and Denoising Algorithm for Phase Imaging

1: **Input:** Wrapped phase image φ and the threshold parameter λ.
2: **Method:**
3: $\rho_{i,j} = \left(\Delta\varphi^x_{i,j} - \Delta\varphi^x_{i-1,j}\right) + \left(\Delta\varphi^y_{i,j} - \Delta\varphi^y_{i,j-1}\right)$.
4: $\hat{\rho} = \mathscr{T}(\rho)$
5: $\hat{\phi}_{i,j} = (\hat{\rho}_{i,j})/2[\cos(\pi i/M) + \cos(\pi j/N) - 2]$
6: $\hat{\phi}_{i,j} = \Theta^{\lambda}_{hard}(\hat{\phi}_{i,j})$
7: **Output:** Restored phase $\phi = \mathscr{T}^{-1}(\hat{\phi})$

5.7.4. Computational complexity. One of the main advantages of the proposed method is its low computational complexity since it is a non-iterative algorithm. Following the SPUD algorithm steps, it can be observed that the DCT transform has a computational complexity of $\mathscr{O}(N\log N)$ and the hard-threshold of $\mathscr{O}(N)$. Therefore, the SPUD algorithm has a computational complexity of $\mathscr{O}(N\log N)$.

6. Simulation Results

This chapter presents the simulation results of the different approaches for the super-resolution scenario and the recovery method.

Metrics: To quantity show the improvement of the proposed methods, the following metric is considered

Relative Error: is computed as the

$$\text{relative error} := \frac{\text{dist}(\mathbf{z},\mathbf{x})}{\|\mathbf{x}\|_2} \tag{130}$$

with

$$dist(\mathbf{x},\mathbf{z}) = \min_{\theta\in[0,2\pi)} \|\mathbf{x}e^{-j\theta} - \mathbf{z}\|_2. \tag{131}$$

where $\mathbf{x}$ is the underlying signal and $\mathbf{z}$ is the estimated.

Peak signal to noise ratio (PSNR): The PSNR between the reference image $\mathbf{x}$ and its estimation $\mathbf{z}$ are calculated mathematically as follows

$$PSNR = 10\log_{10} \frac{||\mathbf{x}||_\infty n}{||\mathbf{x}-\mathbf{z}||_2^2}, \tag{132}$$

where n is the dimension of the signal.

Empirical success rate: A trial is declared successful if the relative error is smaller than 10^{-5}.

Standard deviation This is used for phase images error, which is given by Montresor and Picart (2016),

$$\sigma_\varepsilon = \sqrt{\mathbf{E}[\varepsilon^2] - \mathbf{E}[\varepsilon]^2} \ , \tag{133}$$

where $\varepsilon = \mathbf{x} - \mathbf{z}$ is the phase difference between the simulated true phase $\mathbf{x}$, and the restored phase map $\mathbf{z}$. $\mathbf{E}[\cdot]$ denotes the expected value.

Quality Index: models the phase degradation as structural distortions instead of errors Wang et al. (2002). The Quality Index is defined as,

$$Q_{index} = \frac{\sigma_{sr}}{\sigma_s \sigma_d} \cdot \frac{2\mu_s \mu_r}{\mu_s^2 + \mu_r^2} \cdot \frac{2\sigma_s \sigma_r}{\sigma_s^2 + \sigma_r^2} \ , \tag{134}$$

where μ_s and μ_r denote the mean values of the true simulated phase and the restored phase map, respectively. σ_s and σ_r are their variances and σ_{sr} the covariance. The value of Q_{index} is defined to lie in the interval $[-1, 1]$, being 1 a perfect similarity.

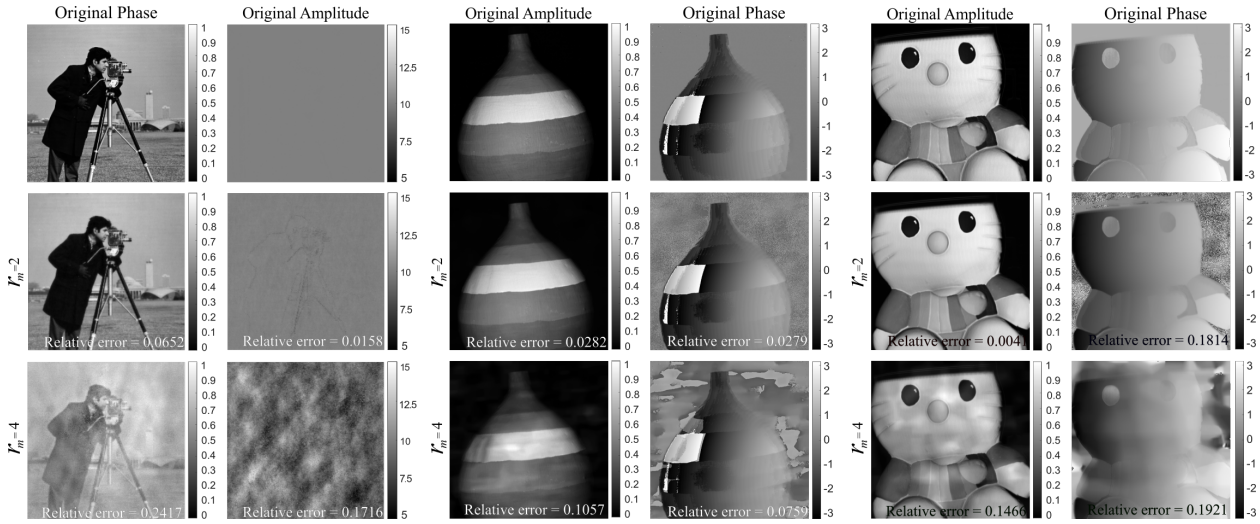
6.1. Physical super resolution phase retrieval

6.1.1. *Super Resolution Factor*. This section illustrates the reconstruction quality attained with the proposed super-resolution scenario described in Section 3.1 using the proposed recovery method described in Section 5.5 for two different super-resolution factors, $r_m = 2, 4$, when the number of experiments/snapshot is fixed to $L = 4$. The coded diffraction patterns at the far zone were simulated using the admissible random variable $d = -1, 1, -j, j$, and the results obtained are shown in Figure 11. For all the images, it can be seen that when the resolution factor increases, the quality of the images decreases. This is an expected result since the diffraction patterns of the

scene are obtained with a detector of poor resolution. Specifically, the reconstructed phase and magnitude details are lost, yielding 0.1466 and 0.1921 relative errors for the magnitude and phase, respectively. This shows that a high-resolution image can be reconstructed using low-resolution sensors, with $r_m < 4$.

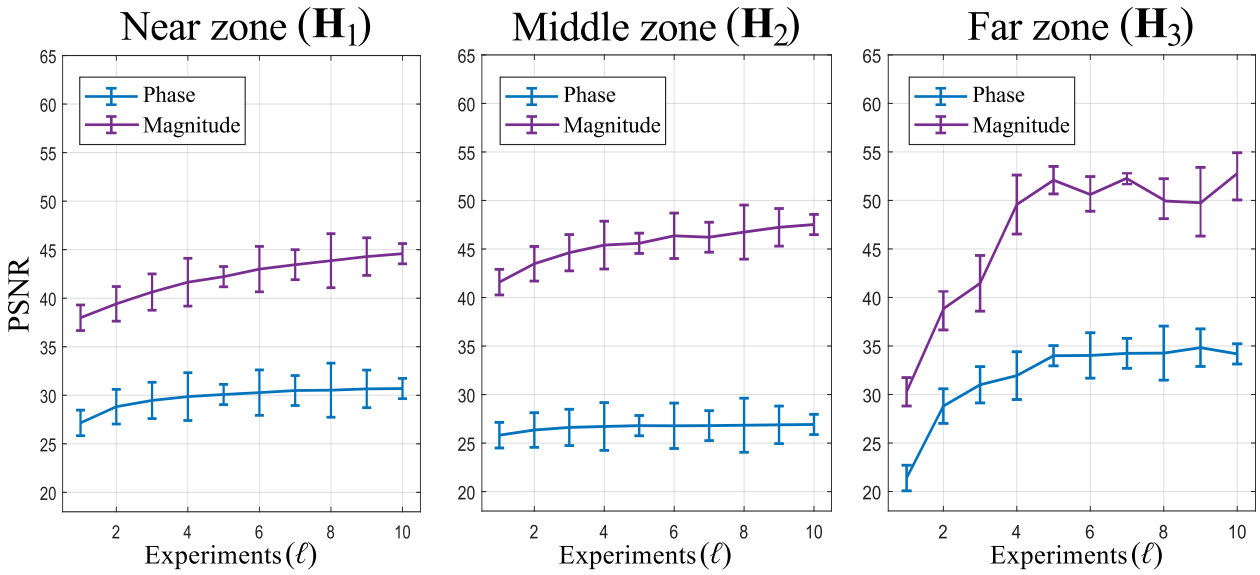
6.1.2. Sampling and Time Complexity. To evaluate the sampling complexity of the proposed method, some numerical tests varying the number of experiments were conducted for a noiseless scenario. Specifically, L is ranged from 1 to 10, using designed, coded apertures based on the random variable d_3. Figure 12 plots the mean and standard deviation of the PSNR of 10 trials. From Fig. 12 it can be observed that when $L \geq 4$, the quality of the reconstruction for both phase and magnitude does not significantly increase. Also, it can be concluded that the reconstruction quality of the magnitude exhibits a higher standard deviation compared to the phase.

Figure 11. Reconstructed images using the proposed method for three different data sets.



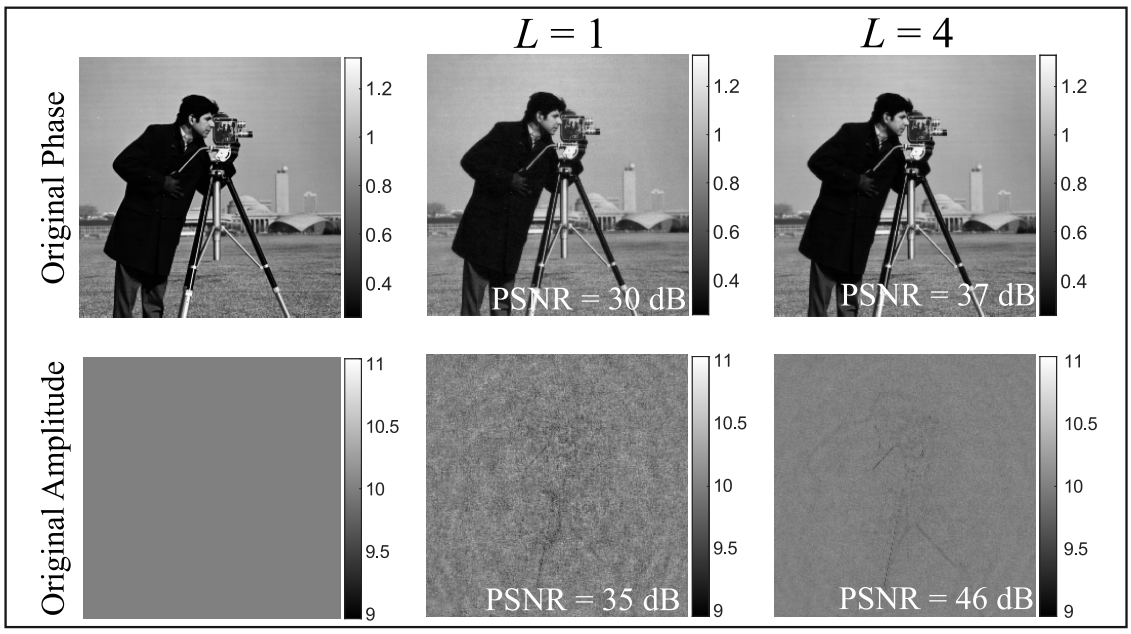
Note: The super-resolution factor is $r_m = 2$ and 4.

Figure 12. Quality of the reconstructed phase and magnitude measured in PSNR



Note: For different number of experiments.

Figure 13. Reconstructed images when $L = 1$ and $L = 4$.



Note: (Top) original and reconstructed phase and (Bottom) amplitude images.

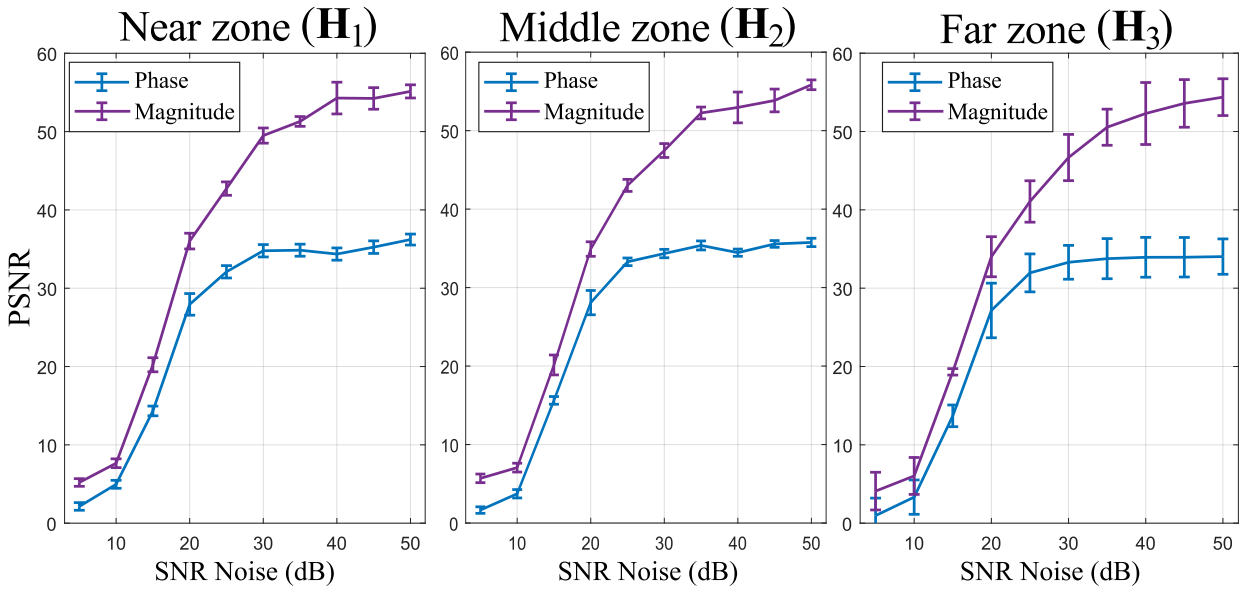
To complement this experiment, Fig. 13 illustrates the reconstructed magnitude and phase using $L = 1,4$ from CDP acquired at the middle zone. Notice that the proposed method obtained 30 dB and 37 dB of PSNR in the reconstructed phase for $L = 1$ and $L = 4$, respectively. For the amplitude 35 dB and 46 dB of PSNR were respectively obtained with the same number of experiments. On the other hand, the running time of Algorithm 7 directly depends on its alternate steps, which for this experiment correspond to $\simeq 15$ sec for Algorithm 8 and $\simeq 0{,}1$ sec for Algorithm 9, 10, implying a total of $\simeq 600$ sec for Algorithm 7.

6.1.3. ***Noise Robustness*.** This section presents numerical simulations to characterize the robustness of the proposed method when the measurements are corrupted by additive Gaussian noise for different values of SNR and for the three diffraction zones. Figure 14 plots the attained Peak-Signal-to-Noise-Ratio (PSNR) with the proposed method using $L = 4$ and $r_m = 2$, and coded apertures based on the random variable d_3, when the SNR is varied from 5 to 50 dB.

Figure. 14 reveals that the reconstruction quality of the proposed method is more stable from 30 to 50 dB of SNR for both magnitude and phase. Further, from values between 5 and 20 dB of SNR, the quality of the reconstruction is low due to the amount of added noise. This shows the effectiveness of the proposed method to recover a high-resolution signal from low-resolution coded diffraction patterns under additive Gaussian noise.

6.1.4. ***Comparison with Other Super-Resolution Schemes*.** From Table 1 it can be observed that the proposed SR models are different from those in Katkovnik et al. (2017); Katkovnik and Egiazarian (2017); Jaganathan et al. (2016). Therefore, to validate the performance of the algorithm concerning other SR schemes, Algorithm 7 was adapted to the sensing model at the

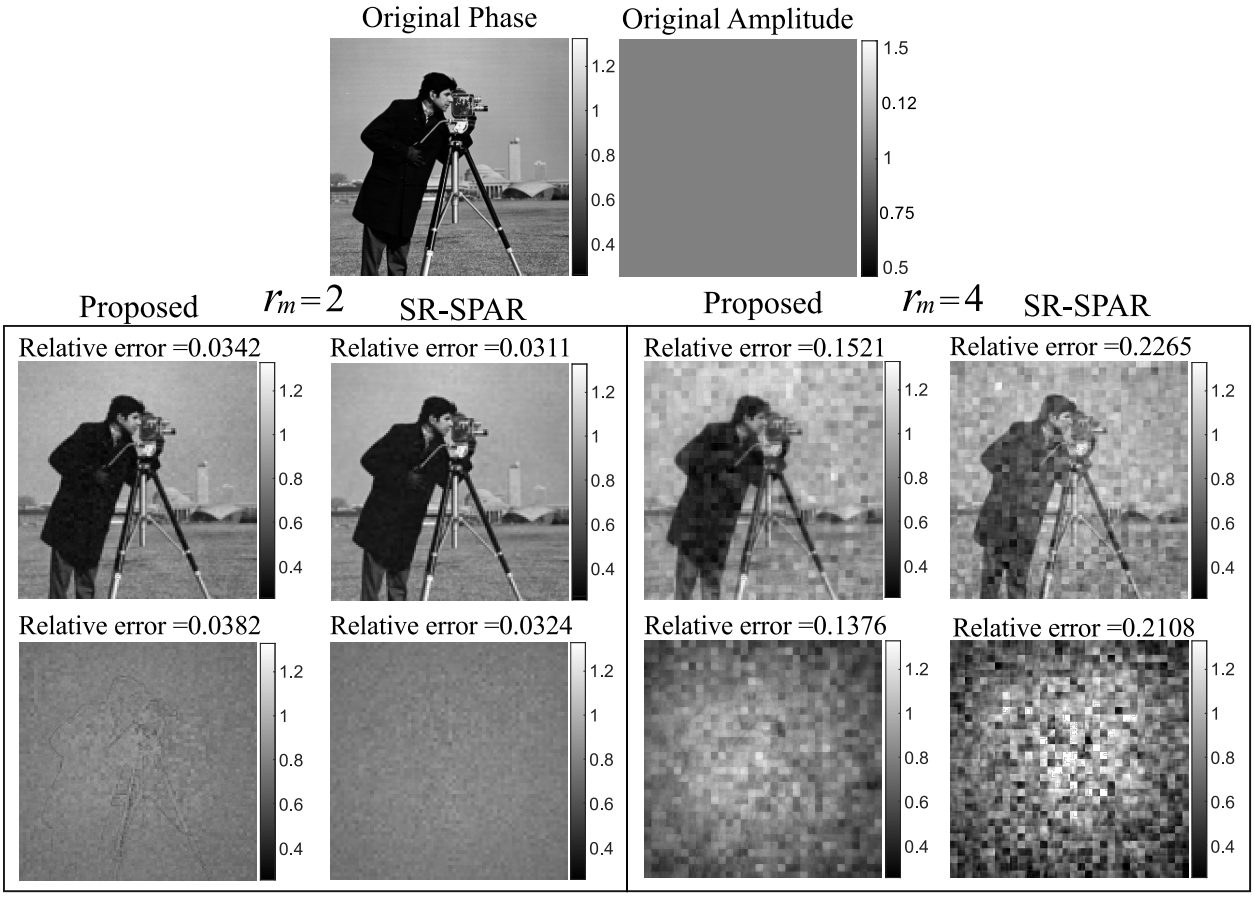
Figure 14. Reconstructed quality of the phase and magnitude measured in PSNR.



Note: For different levels of noise.

far-field in Katkovnik and Egiazarian (2017) and compared with its computational super-resolution algorithm, SR-SPAR. Further, in this section, this thesis does not compare with the methodology proposed in Jaganathan et al. (2016) because its reconstruction algorithm becomes intractable for large images, as in the case of the used images. This test was carried out without noise for $L = 2$, using a set of coded apertures based on the random variable d_3. All the parameters used for SR-SPAR were those suggested in Katkovnik and Egiazarian (2017). The pixel sizes of the sensor and the coded aperture were fixed as $\Delta_s = \Delta_m = 5{,}2\mu m$, and the wavelength λ is assumed to be $\lambda = 632{,}8nm$. The number of pixels of the simulated sensor is 1024×1024, assuming that it fully covers the diffraction pattern area. Two super-resolution factors are compared, $r_m = 2, 4$, and the attained results of the proposed methods and SR-SPAR are shown in Fig. 15. Observe that these results suggest that the proposed reconstruction method exhibits a comparable performance with

Figure 15. Comparison with state-of-the-art



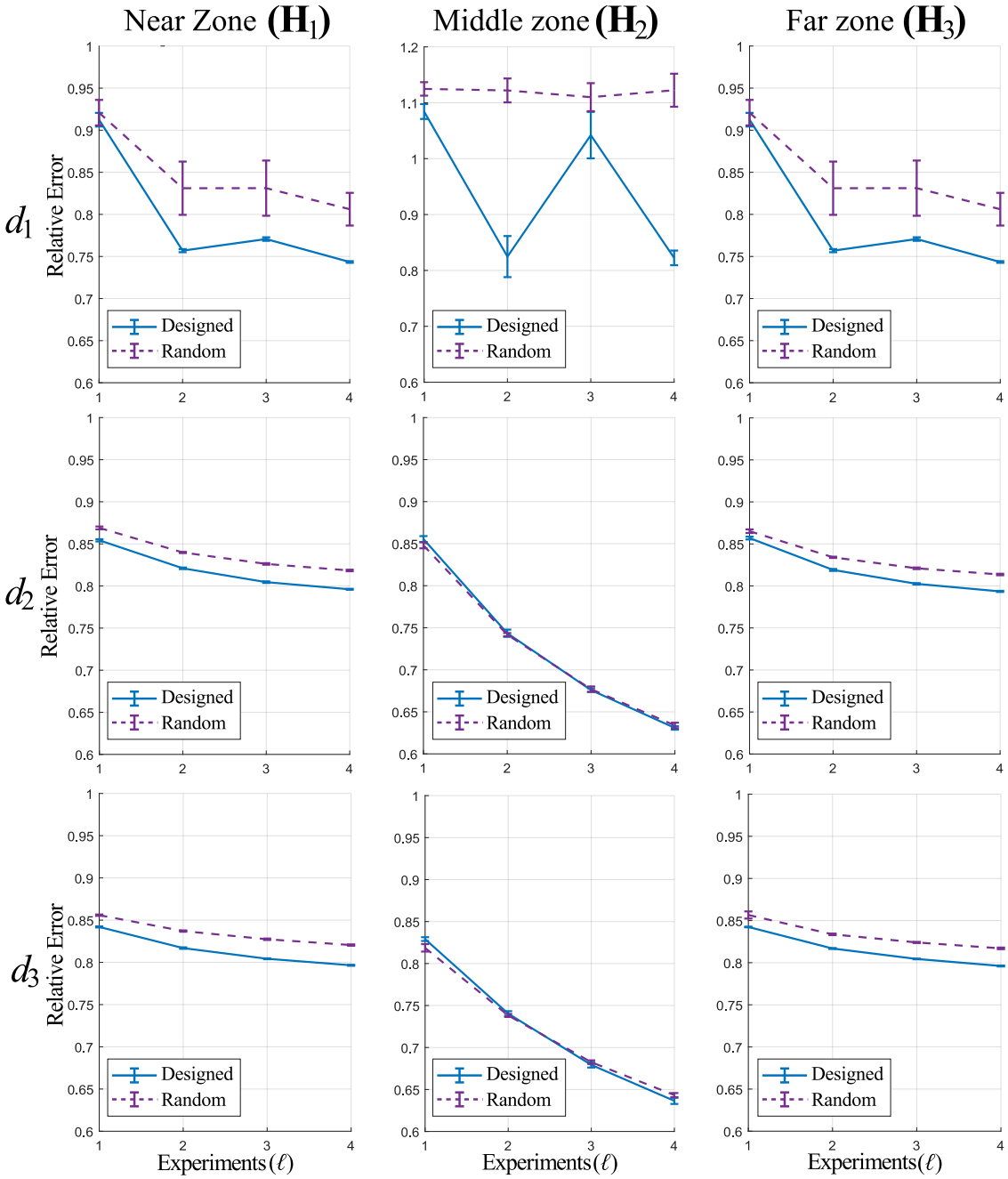
Note: (Top) Original phase and amplitude, (Middle) reconstructed phase for $r_m = 2$ and $r_m = 4$ and (Bottom) reconstructed magnitude for the same super-resolution factors.

respect to SR-SPAR under the super-resolution scenario in Katkovnik and Egiazarian (2017).

6.2. Coded Aperture Design for Super-Resolution

As explained in section 5, the initialization plays an important role in correcting phase image estimation. Therefore the performance of the well-known orthogonal-promoting initialization proposed in Wang et al. (2017b) is evaluated. Therefore, the designed coded apertures, using uniform random variables $d_1 \in \{0,1\}, d_2 \in \{-1,1\}$ and $d_3 \in \{-1,1,-j,j\}$ is compared with random

Figure 16. Relative error of the returned initialization using designed and non-designed coded apertures



Note: The number of experiments is varied.

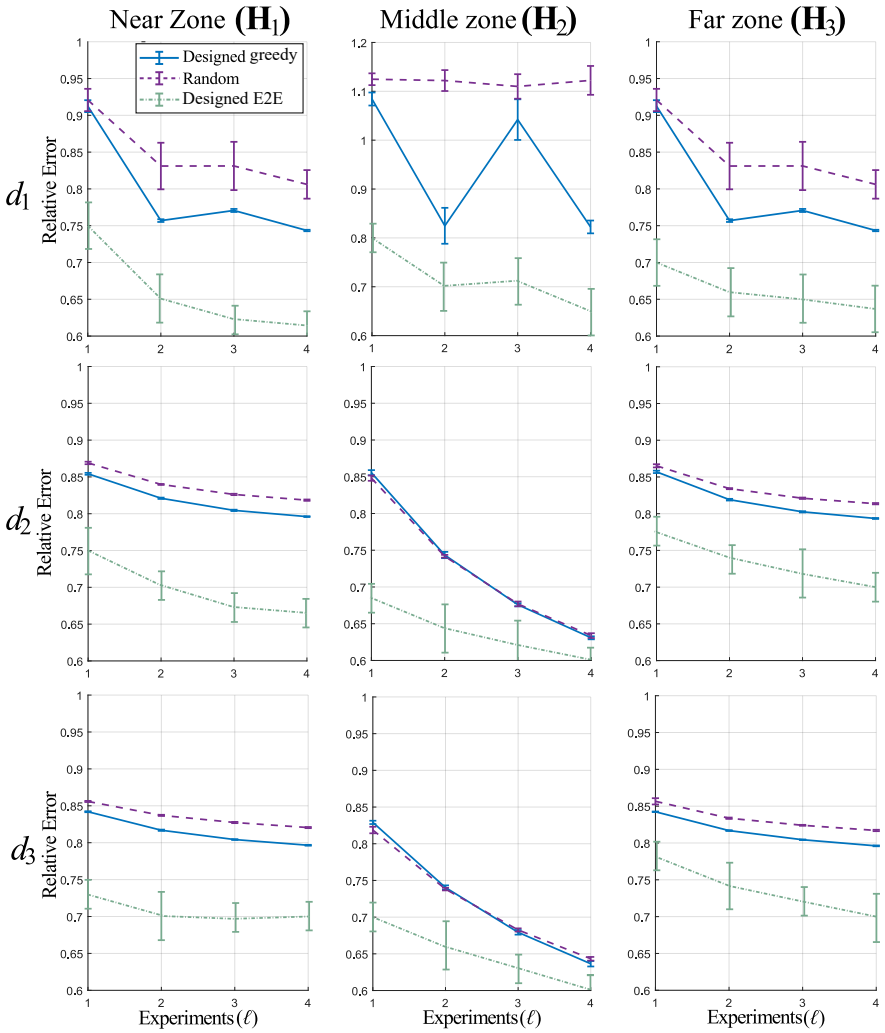
distribution. Specifically, the relative error of the returned initialization for different number of experiments, and the constant δ in (46) were determined for these random variables. The results are summarized in Fig. 16 and Table 2. From Fig. 16 it can be observed that the designed coded apertures generate a more accurate initialization of the true image compared with non-designed ensembles for any diffraction zone and for all d_1, d_2 and d_3. These numerical results are expected since the designed coded apertures are constructed to satisfy the condition better (44). Additionally, to verify that the designed coded apertures could lead to better estimations of the true signal, the minimum constant δ from (46) is determined and presented in Table 2, when $L = 4$. Specifically, observe that the attained value of δ when designed coded apertures are used is smaller than that of non-designed ensembles. Remark that (46) states that a small value of δ is desired, which is independent of the diffraction zone according to condition (44). In summary, these numerical tests validate the proposed coded aperture design strategy.

Table 2. Value of δ using designed coded apertures for random variables d_1, d_2 and d_3, when $L = 4$.

δ	d_1	d_2	d_3
Proposed	**0.3750**	**0.5807**	**0.5984**
Random	0.5568	0.6564	0.6167

6.2.1. ***E2E Phase Mask Design.*** The End-to-End (E2E) approach considers a data-set to find the spatial distribution of the CA by customizing the optical sensing as a layer in an E2E network. For this, the NYU Depth Dataset Silberman et al. (2012) is employed for evaluating the proposed deep approach, which contains 1449 RGB images with depth maps of 15 discretization levels; here, 80 %, 10 %, and 10 % of images were selected for training, testing, and validating, respectively. On the one hand, all images were resized to 256×256 pixels; then, the RGB images

Figure 17. Relative error of the returned initialization using designed and non-designed coded apertures.



Note: The number of experiments is varied.

were converted to a grayscale version and normalized to simulate the amplitude information. On the other hand, the depth maps were scaled in the range $[-\pi, \pi]$ to simulated the phase information. Similar as the previous section, this thesis evaluated the designed coded apertures, using random variables $d_1 \in \{0,1\}, d_2 \in \{-1,1\}$ and $d_3 \in \{-1,1,-j,j\}$ compared with random distribution and

the previous design. To address the constraints of the variables, this thesis starts with random values between $e^{j[-\pi,\pi]}$ and the following regularize integrated into the main loss function of the network

$$R(D = e^{j\mathbf{C}}) = \frac{1}{L}\sum_{i,j,\ell}\prod_{d}\left(\mathbf{C}_{i,j}^{\ell} - \kappa_d\right)^2, \tag{135}$$

where k_d is the desired value. For instance, for $\{-1,1\}$ the $k_1 = \pi$ and $k_2 = \pi$ which produce $e^{j\pi} = 1$ and $e^{-j\pi} = -1$. For this experiment, the network is composed of the forward model and also the initialization, as is illustrated in Fig.9. The MSE is used as the main loss function. The relative error of the returned initialization for the different numbers of experiments is summarized in Fig.17. It can be observed that the designed, coded apertures generated by the E2E method significantly improve the results obtained with the Random and the greedy strategy presented in the previous section.

6.3. Recovery Phase Retrieval Algorithms

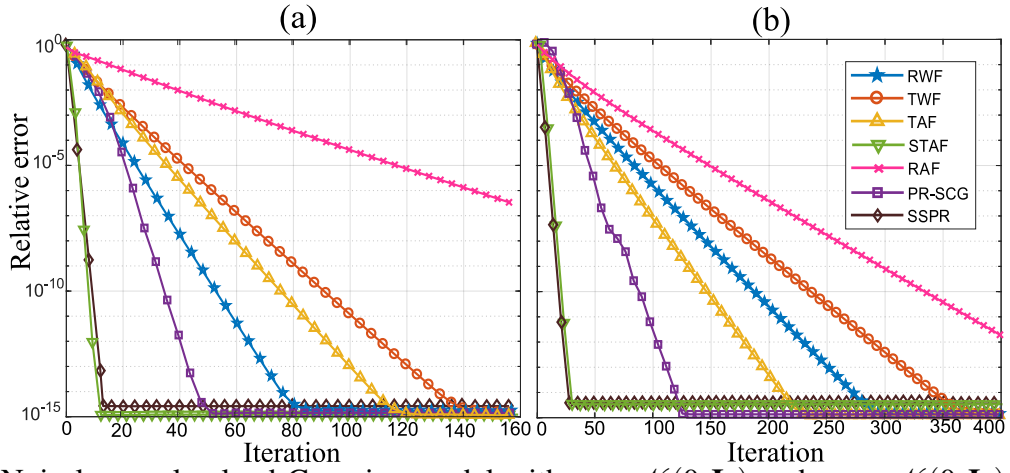
6.3.1. *Smoothing and Stochastic Smoothing Phase Retrieval Algorithm.*

6.3.1.1. Sampling Complexity and Speed of Convergence. For this experiment, simulations compare the convergence speed and sample complexity when all the algorithms are equipped with their initialization and suggested parameter settings for noiseless real and complex data. In the case of the incremental algorithms such as SSPR and STAF, one iteration is equivalent to a m stochastic iterations over the entire data, *i.e.* m gradient evaluations of the component functions ℓ_{k_t}. Figure 18 summarizes the converging speed of the different mentioned methods.

The real case scenario results are shown in Fig. 18 (a). Note that PR-SCG requires up to

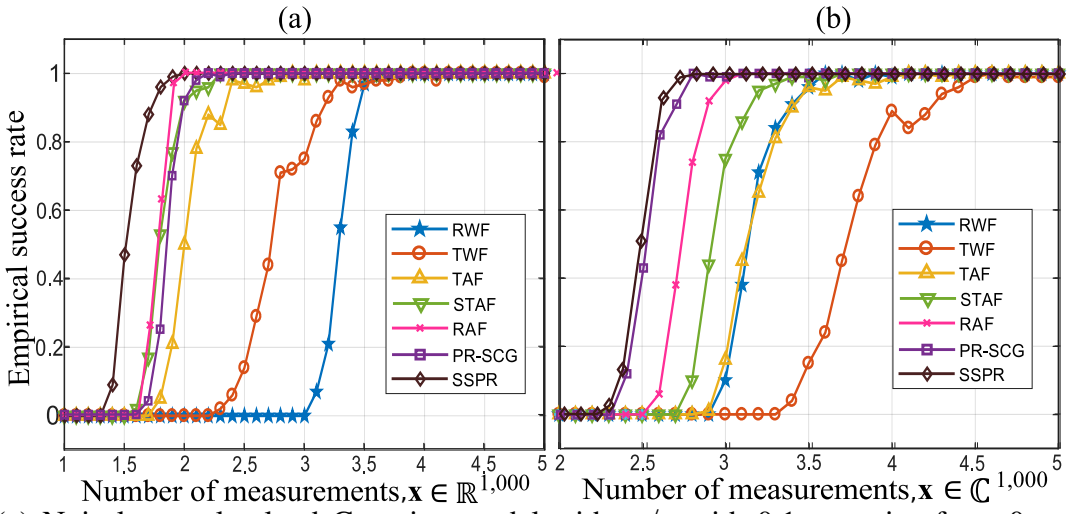
66.7 %, 58.3 %, and 41.1 % fewer number of iterations concerning TWF, TAF, and RWF algorithms, respectively, and SSPR can solve the phase retrieval problem with 89.5 %, 87.2 %, 82.1 % and 69.4 % less number of iterations in contrast to Truncated Wirtinger Flow (TWF) Chen and Candes (2015a), Truncated Amplitude Flow (TAF) Wang et al. (2016a), Reshaped Wirtinger Flow (RWF) Zhang and Liang (2016) and the proposed PR-SCG algorithms respectively. Note that Stochastic Truncated Amplitude Flow (STAF) Wang et al. (2017a) and the proposed SSPR require a similar number of iterations for both real and complex cases as shown in Figs. 18(a) and 18(b), respectively. Further, for the complex scenario, shown in Fig. 18(b), the PR-SCG method can solve the phase retrieval problem with 28.5 %, 45.6 % and 67.1 % less number of iterations in contrast to TAF, RWF, and TWF, respectively, and SSPR requires up to 79.2 %, 88.8 %, 90.7 %, and 93.1 % less number of iterations with respect to PR-SCG, RWF, TAF, and TWF algorithms, respectively,

Figure 18. Relative error versus iteration with $n = 1{,}000$, $m/n = 8$.



Note:(a) Noiseless real-valued Gaussian model with $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_n)$ and $\mathbf{a}_k \sim \mathcal{N}(0, \mathbf{I}_n)$, (b) Noiseless complex-valued Gaussian model with $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_n) + j\mathcal{N}(0, \mathbf{I}_n)$ and $\mathbf{a}_k \sim \mathcal{N}(0, \frac{1}{2}\mathbf{I}_n) + j\mathcal{N}(0, \frac{1}{2}\mathbf{I}_n)$.

Figure 19. Empirical success rate versus number of measurements with $n = 1{,}000$, m/n.



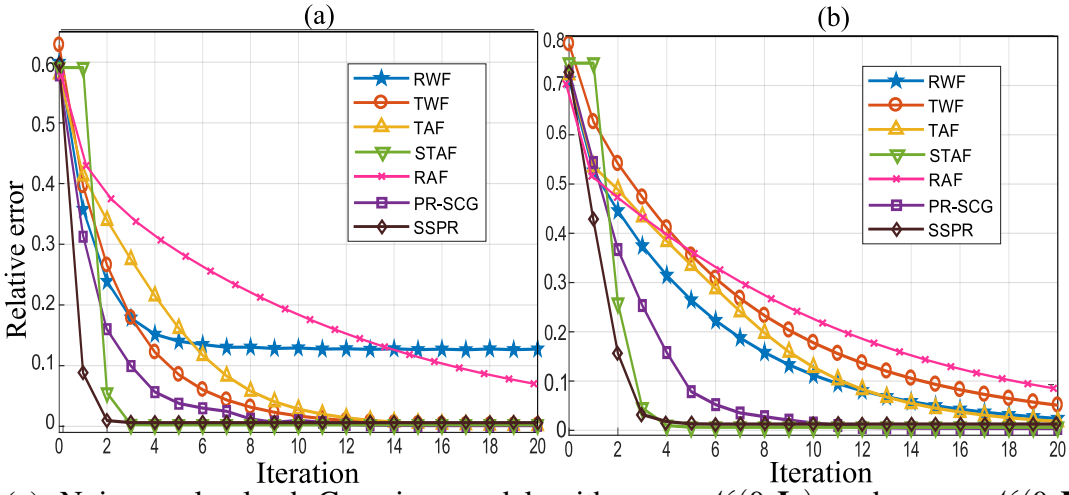
Note: (a) Noiseless real-valued Gaussian model with m/n with 0.1 step size from 0 to 7, $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_n)$ and $\mathbf{a}_k \sim \mathcal{N}(0, \mathbf{I}_n)$. (b) Noiseless complex-valued Gaussian model, m/n with 0.1 step size, $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_n) + j\mathcal{N}(0, \mathbf{I}_n)$ and $\mathbf{a}_k \sim \mathcal{N}(0, \frac{1}{2}\mathbf{I}_n) + j\mathcal{N}(0, \frac{1}{2}\mathbf{I}_n)$.

and similar iterations in contrast to STAF. Notice that the RAF method requires more iterations to converge than the other analyzed algorithms for both complex and real cases.

Additionally, numerical results were conducted to validate the sample complexity of the different algorithms for real and complex noiseless cases. In particular, the results for the real case, shown in Fig. 19(a), establishes that SSPR achieves a success rate of over 93 % when $m/n = 1{,}8$ and guarantees perfect recovery from about $1{,}9n$ measurements. For the case of PR-SCG, a success rate of around 98 % is obtained when $m/n = 2{,}1$ and perfect recovery from about $2{,}2n$ measurements.

Now, considering the complex case, Fig. 19(b) shows that the PR-SCG algorithm ensures perfect recovery from about $2{,}8n$ measurements and SSPR achieves perfect recovery from about $2{,}7n$ measurements. Therefore, these numerical results confirm the effectiveness of the proposed methods.

Figure 20. Relative error versus iteration with $n = 1,000$ and $m/n = 8$.

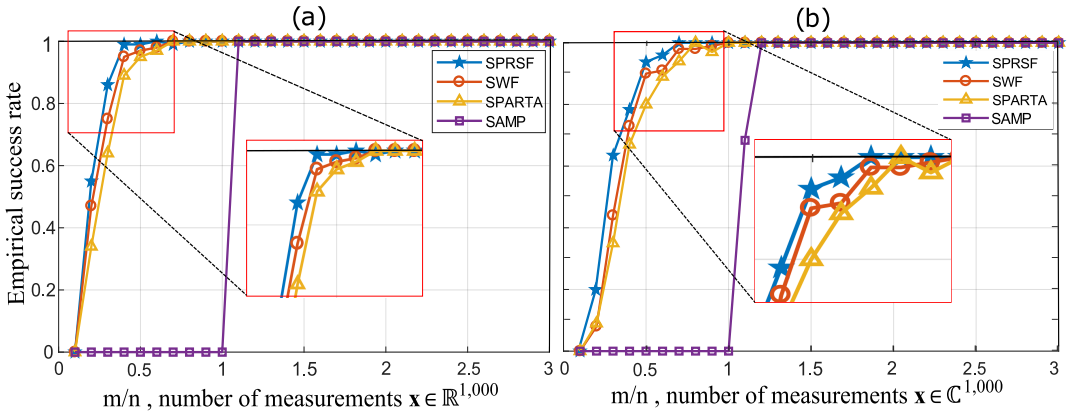


Note: (a) Noisy real-valued Gaussian model with $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_n)$ and $\mathbf{a}_k \sim \mathcal{N}(0, \mathbf{I}_n)$. (b) Noisy complex-valued Gaussian model with $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_n) + j\mathcal{N}(0, \mathbf{I}_n)$ and $\mathbf{a}_k \sim \mathcal{N}(0, \frac{1}{2}\mathbf{I}_n) + j\mathcal{N}(0, \frac{1}{2}\mathbf{I}_n)$.

6.3.1.2. Noise Robustness. Numerical tests are conducted to demonstrate the robustness of PR-SCG and SSPR to additive noise corruption. These simulations were realized under the noisy real and complex-valued Gaussian model $\hat{y}_k = |\mathbf{a}_k^H \mathbf{x}|^2 + \eta_k$ with $\eta_k \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$. The noisy data was generated as $q_k = \sqrt{\hat{y}_k}$ and $\sigma^2 = 0{,}1^2 \|\mathbf{x}\|_2^2$ with $m/n = 8$. Figure 20 summarizes these experiments for the real and complex cases. Specifically, it can be observed that the SSPR exceeds in convergence speed of the other methods in real and complex cases since it requires fewer iterations to solve the phase retrieval problem. In terms of the number of iterations, STAF exhibits the same behavior as SSPR according to the previous test, but under a noisy scenario, the performance of STAF diminishes compared to SSPR. Thus, these experiments show a better statistical performance of the SSPR method under a noise corruption model.

Figure. 20 also shows that PR-SCG attains a higher performance under a noisy scenario

Figure 21. Empirical success rate versus number of measurements for $n = 1000$, known sparsity $k = 10$ and m/n with a step size of 0.1 from 0.1 to 3.



Note: (a) Noiseless real-valued Gaussian model for $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_{1000})$ and $\mathbf{a}_i \sim \mathcal{N}(0, \mathbf{I}_{1000})$. (b) Noiseless complex-valued Gaussian model, with $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_n) + j\mathcal{N}(0, \mathbf{I}_n)$ and $\mathbf{a}_i \sim \mathcal{N}(0, \frac{1}{2}\mathbf{I}_n) + j\mathcal{N}(0, \frac{1}{2}\mathbf{I}_n)$.

than its non-stochastic competing alternatives TAF, RWF, and TWF.

6.3.2. *Sparse Smoothing Phase Retrieval Problem.*

6.3.2.1. Known Sparsity. The first experiment analyzes the sampling complexity under a noiseless real and complex Gaussian model, assuming that the sparsity k is known. Figure 21 summarizes the attained empirical success rate in terms of the number of measurements, for all algorithms under analysis. For this test, the sparsity of the signal $\mathbf{x}$ is fixed as $k = 10$, and the ratio between m and n ($i.e\ m/n$) is varied from 0.1 to 3, with a step size of 0.1, for both the real and the complex cases. At each ratio m/n, the average over 100 tests was calculated.

The simulations in Fig. 21 suggest that the proposed algorithm SPRSF requires less number of measurements to solve the sparse phase retrieval problem in comparison with the SparseAltMinPhase (SAMP) Netrapalli et al. (2013), sparse WF (SWF) Yuan et al. (2017), and the Sparse Truncated Amplitude flow (SPARTA) Wang et al. (2016b) methods, for both the real and the com-

plex cases. Moreover, notice that SPRSF achieves a success rate over 98 % when $m/n = 0{,}5$ for the real case and a success rate over 95 % when $m/n = 0{,}6$ for the complex case. Further, SPRSF guarantees a perfect recovery from about $0{,}6n$ and $0{,}7n$ measurements for the real and complex cases, respectively. Therefore, these results show the effectiveness of the smoothing approximation scheme to solve the sparse phase retrieval problem.

6.3.2.2. Unknown Sparsity Boundary. In this experiment, this thesis compares the proposed method with the state-of-the-art approaches to recover the signal $\mathbf{x}$ in terms of the sampling complexity, when the sparsity k is unknown. Specifically, from Theorem 5, it can be obtained that the sampling complexity of the SPRSF method is $\mathscr{O}(k^2 \log(n))$. Now, suppose that there is no knowledge about the sparsity k. It is assumed that the sparsity is $k = \sqrt{n}$, the sampling complexity is given by $\mathscr{O}(n \log(n))$, which is considered the limit value of the unknown k when $k \ll n$ Yuan et al. (2017). Therefore, in this Test the sparsity of the signal $\mathbf{x}$ is fixed to $k = 10$, but the experiments, in Fig. 22, assume the sparsity of the signal $\mathbf{x}$ is $\sqrt{n} \approx 32$, since $n = 1000$.

Notice that, SPRSF outperforms the other algorithms when the prior sparsity k is not known correctly for both real and complex cases. Further, it can be observed that compared with Test 1 in Fig. 21, the superiority of the proposed method SPRSF with respect to SPARTA, SWF and SAMP, is more evident. Figure 22 also shows that SPRSF attains a success rate of 80 % when $m/n = 0{,}3$ for the real case and a success rate of 90 % when $m/n = 0{,}5$ for the complex case. Perfect recovery is attained from about $0{,}6n$ and $0{,}7n$ measurements for the real and the complex cases, respectively.

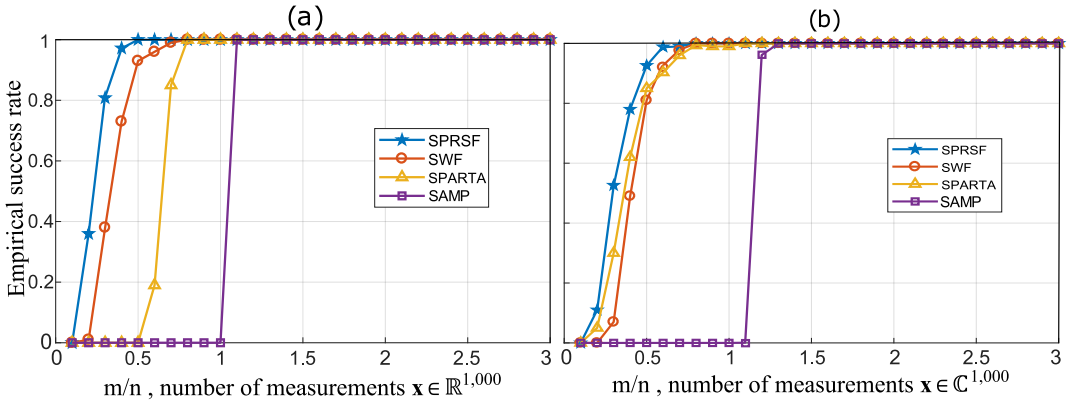
It can be concluded that this second test suggests that the proposed smoothing approximation scheme outperforms its competitive alternatives when the sparsity is assumed different from

its real value.

6.3.2.3. Unknown Sparsity. In this experiment, numerical simulations are conducted to analyze the ability of the methods to solve the sparse phase retrieval problem when the sparsity k is completely unknown. For these simulations, the sparsity of the signal $\mathbf{x}$ was fixed as $k = 10$, and since the sparsity is unknown, this thesis ranges $\hat{k}$ from 35 to 180 for real and complex cases, with a step size of 5. At each $\hat{k}$, the average of the empirical success rate over 100 tests is calculated. This thesis called the sparsity $\hat{k}$, the prior sparsity. The number of measurements m was fixed to $m = n$. All these numerical tests are summarized in Fig. 23. This thesis omitted the SAMP simulations in Fig. 23 since from Fig. 21 it can be noticed that SAMP cannot solve the sparse PR problem when the sparsity k is known and the number of measurements $m = n$.

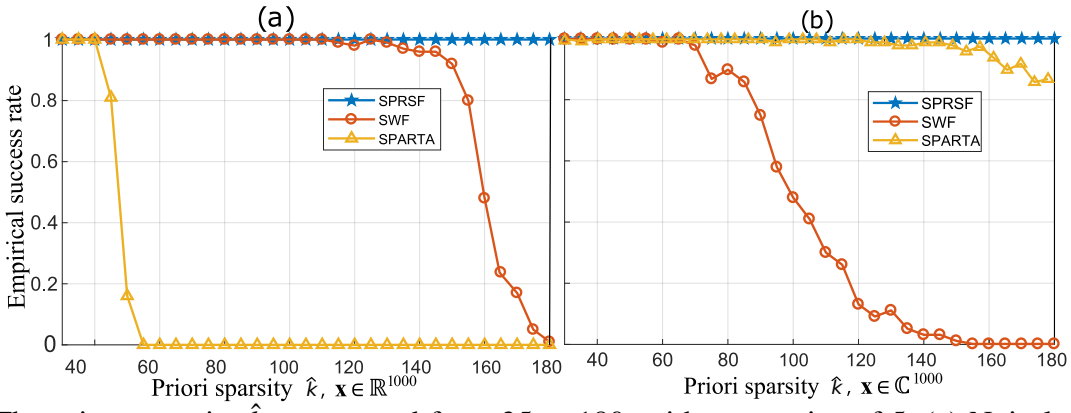
From Fig. 23 it can be observed that the proposed method SPRSF overcomes its competing alternatives because it guarantees perfect recovery when the sparsity k of the signal $\mathbf{x}$ is completely

Figure 22. Empirical success rate versus number of measurements for $n = 1{,}000$, m/n with a step size of 0.1 from 0 to 3.



Note: The sparsity is assumed to be $k = \sqrt{n} \approx 32$ while the real sparsity is $k = 10$. (a) Noiseless real-valued Gaussian model for $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_{1000})$ and $\mathbf{a}_i \sim \mathcal{N}(0, \mathbf{I}_{1000})$. (b) Noiseless complex-valued Gaussian model, with $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_n) + j\mathcal{N}(0, \mathbf{I}_n)$ and $\mathbf{a}_i \sim \mathcal{N}(0, \frac{1}{2}\mathbf{I}_n) + j\mathcal{N}(0, \frac{1}{2}\mathbf{I}_n)$.

Figure 23. *Empirical success rate versus number of measurements for $n = 1000$, $m/n = 1$, where the real sparsity is $k = 10$.*



Note: The priory sparsity $\hat{k}$ was ranged from 35 to 180, with a step size of 5. (a) Noiseless real-valued Gaussian model with $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_{1000})$ and $\mathbf{a}_i \sim \mathcal{N}(0, \mathbf{I}_{1000})$. (b) Noiseless complex-valued Gaussian model with $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_n) + j\mathcal{N}(0, \mathbf{I}_n)$ and $\mathbf{a}_i \sim \mathcal{N}(0, \frac{1}{2}\mathbf{I}_n) + j\mathcal{N}(0, \frac{1}{2}\mathbf{I}_n)$.

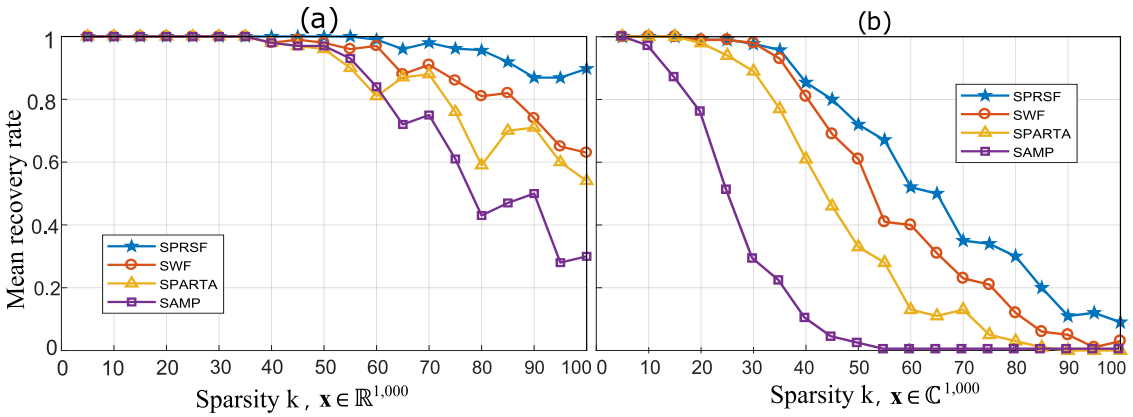
unknown. Further, notice that SPARTA cannot recover the signal without prior knowledge about the sparsity from a prior sparsity $\hat{k} = 55$ and $\hat{k} = 140$ for the real and complex cases, respectively when the sparsity is $k = 10$. Also, it can be concluded that SWF is superior to SPARTA for the real case, but SWF cannot recover the sparse signal from a prior sparsity $\hat{k} \geq 155$. However, for the complex case, SPARTA exhibits a better performance than SWF, because SPARTA cannot always recover the signal from a priority sparsity $\hat{k} \geq 150$.

In summary, by combining the results from Test 2 (Fig. 22) and Test 3 (Fig. 23), it can be concluded that SPRSF is highly superior to SPARTA, SAMP, and SWF in recovering the sparse signal $\mathbf{x}$ when there is no prior knowledge about the sparsity k.

6.3.2.4. Different Values of Sparsity Analysis. This section shows numerical simulations to determine the effect of different sparsity values on the performance of SAMP, SPARTA, SWF and SPRSF. For these experiments, this thesis fixed the number of measurements $m = 1{,}5n$

with $n = 1000$ and the sparsity of the signal varying from 10 to 100 with a step size of 5. In these cases, this thesis assumes that the sparsity k is known. All the numerical results are summarized in Fig. 24.

Figure 24. Empirical success rate versus sparsity k ranged from 10 to 100 with a step size of 5, $n = 1000$, $m/n = 1{,}5$.



Note: (a) Noiseless real-valued Gaussian model for $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_{1000})$ and $\mathbf{a}_i \sim \mathcal{N}(0, \mathbf{I}_{1000})$. (b) Noiseless complex-valued Gaussian model, with $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_n) + j\mathcal{N}(0, \mathbf{I}_n)$ and $\mathbf{a}_i \sim \mathcal{N}(0, \frac{1}{2}\mathbf{I}_n) + j\mathcal{N}(0, \frac{1}{2}\mathbf{I}_n)$.

Figure 24 shows that the SPRSF method is superior to the SAMP, SPARTA and SWF algorithms, for both real and complex cases, since SPRSF can solve the sparse phase retrieval problem for signals with larger sparsity values, as opposed to its competitive alternatives. Also, it can be concluded that SPRSF has a mean recovery rate of about 75 % and 12 % when the sparsity is $k = 100$ for the real and complex cases, respectively.

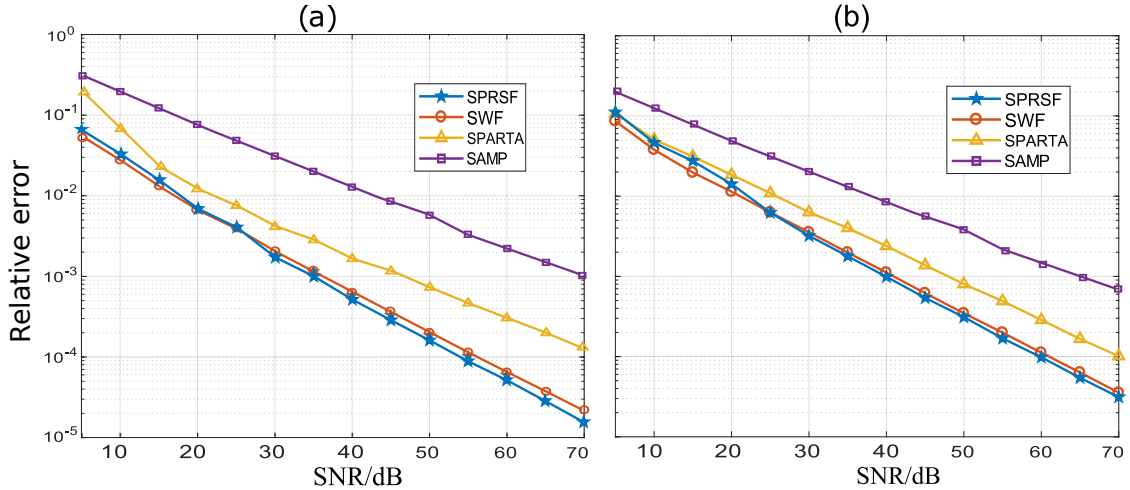
6.3.2.5. Noise Corruption Analysis. Numerical tests are conducted to demonstrate the robustness of SPRSF to noise corruption. These simulations are performed under the noisy real complex valued Gaussian model $\hat{y}_i = |\mathbf{a}_i^H \mathbf{x}| + \eta_i$. The noisy data was generated as $q_i = \hat{y}_i$ with a signal to noise ratio (SNR) ranging from 5dB to 70dB. The number of measurements was fixed

as $m = 1{,}5n$ and the sparsity as $k = 10$. The results in Fig. 25 are the average of the relative error metric $d_r(\mathbf{z},\mathbf{x})/\|\mathbf{x}\|_2$ of 100 tests for each SNR value.

From Fig. 25 it can be observed that SWF attains slightly better performance in solving the sparse phase-retrieval problem, compared with SPRSF for the real and complex cases, in high-noise scenarios $0 < SNR \leq 20$. However, when the noise level decreases, the proposed method overcomes that of SWF for both cases. Further, for the real and complex cases, the results show that SPRSF exhibits better performance than its competitive SPARTA and SAMP alternatives for all noise values.

6.3.3. *Smoothing Phase Retrieval with Outliers*. This section evaluates the numerical performance of the proposed smoothing phase retrieval algorithm with outliers algorithm compared with the competitive algorithms Median-RWF Zhang et al. (2018), Median-TWF Zhang et al. (2018), Robust-RWF Chen et al. (2017), and Prox Duchi and Ruan (2017). Two cases of

Figure 25. Mean of 100 NMSE test for different values of Gaussian white noise from 5dB to 70dB of SNR.



Note: (a) Noisy real-valued Gaussian model. (b) Noisy complex-valued Gaussian model.

outliers were tested:

- **Case 1:** When the measurement vectors and the outliers are independent, i.e.

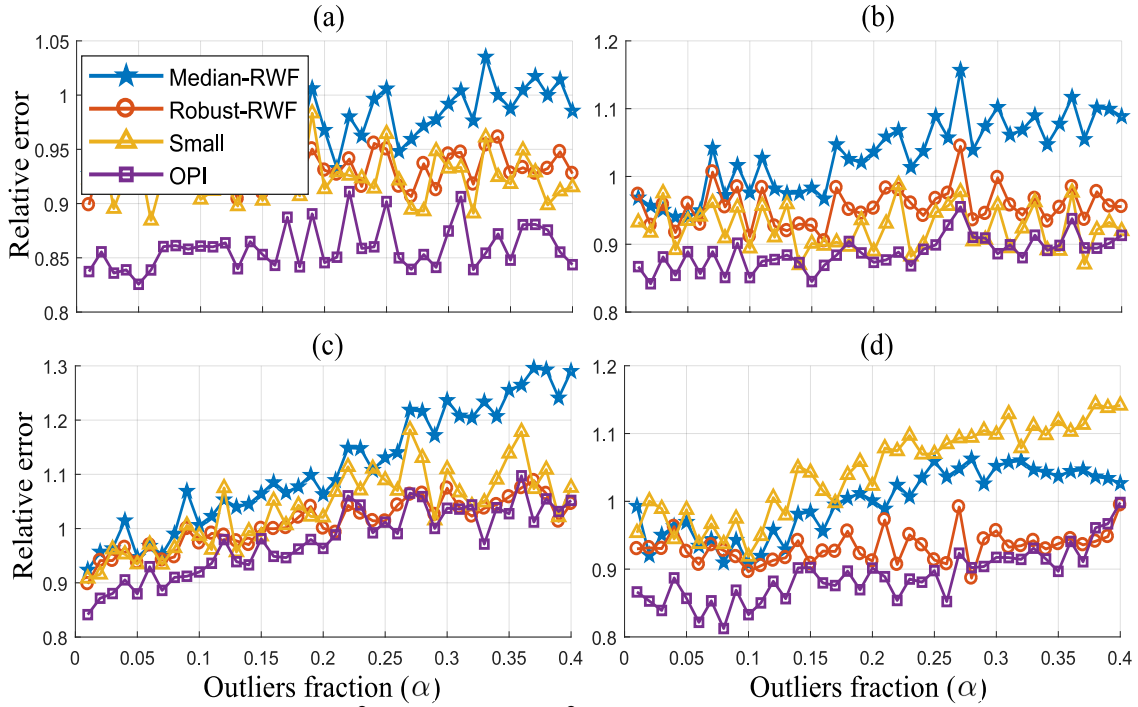
$$y_k = \begin{cases} |\mathbf{a}_k^H \mathbf{x}|^2 & \text{if} \quad \overline{\mathscr{T}^*} \\ o_k, & \text{if} \quad \mathscr{T}^*, \end{cases} \tag{136}$$

where $\mathscr{T}^*$ is the true support of the outliers and $\overline{\mathscr{T}^*}$ is its complement.

- **Case 2:** The corrupted measurements are dependent of the sensing vectors a_k as in (90).

The default values of the parameters of Algorithms 6, were fixed to $\mu_0 = 60, T = 300, \beta =$

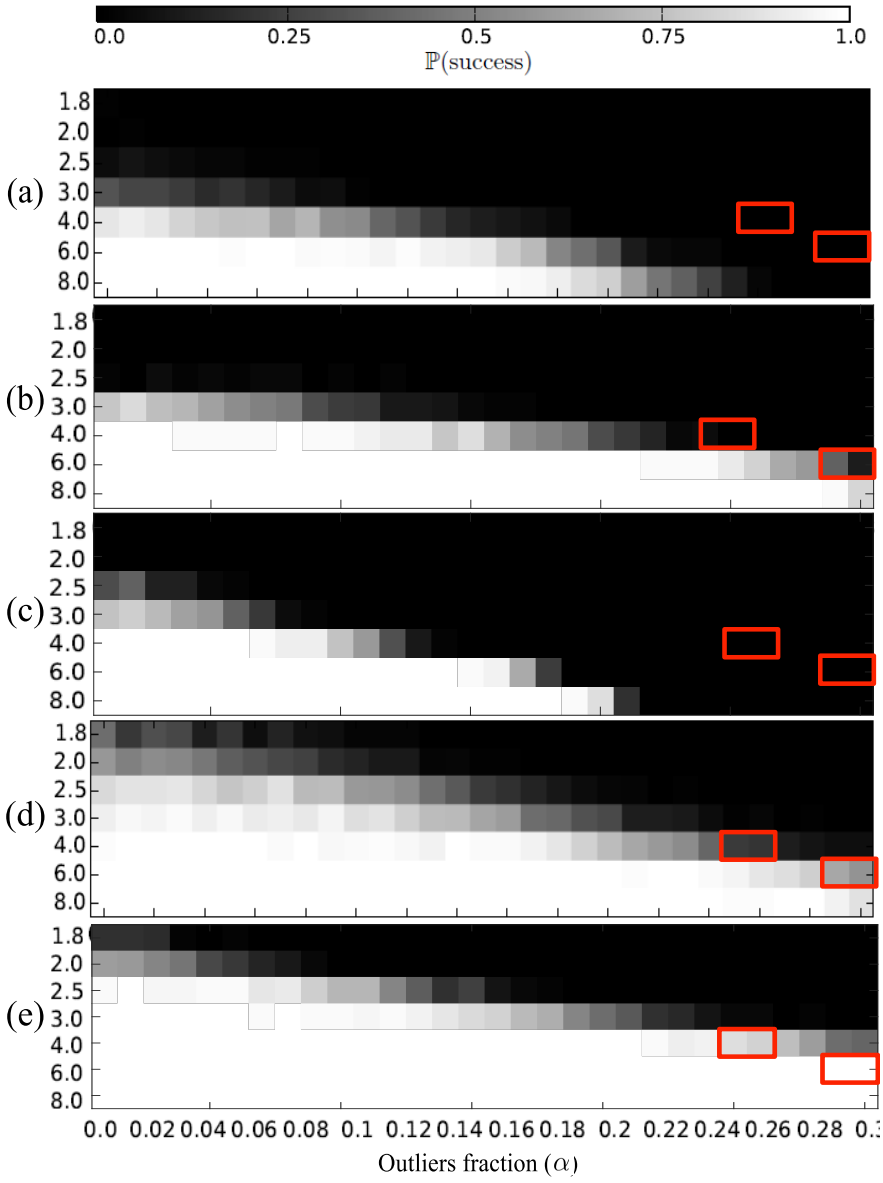
Figure 26. Relative error of the returned initialization



Note: for case (a) $o_{max} = 0{,}5\|\mathbf{x}\|^2$, (b) $o_{max} = \|\mathbf{x}\|^2$, (c) $o_{max} = 0\|\mathbf{x}\|_\infty \|\mathbf{x}\|_2$ and for case 1 (d) setting y_k as 0.

4,6, $\gamma = 0{,}9$, $\gamma_1 = 0{,}5$, $\lambda = 0{,}6$. These values were determined using a cross-validation strategy such that each simulation uses the value that results in the best reconstruction quality.

Figure 27. Probability of success for a) median-TWF b) median- RWF c) Robust-RWF d) Prox and e) the proposed RSPR where dimension $n = 100$.



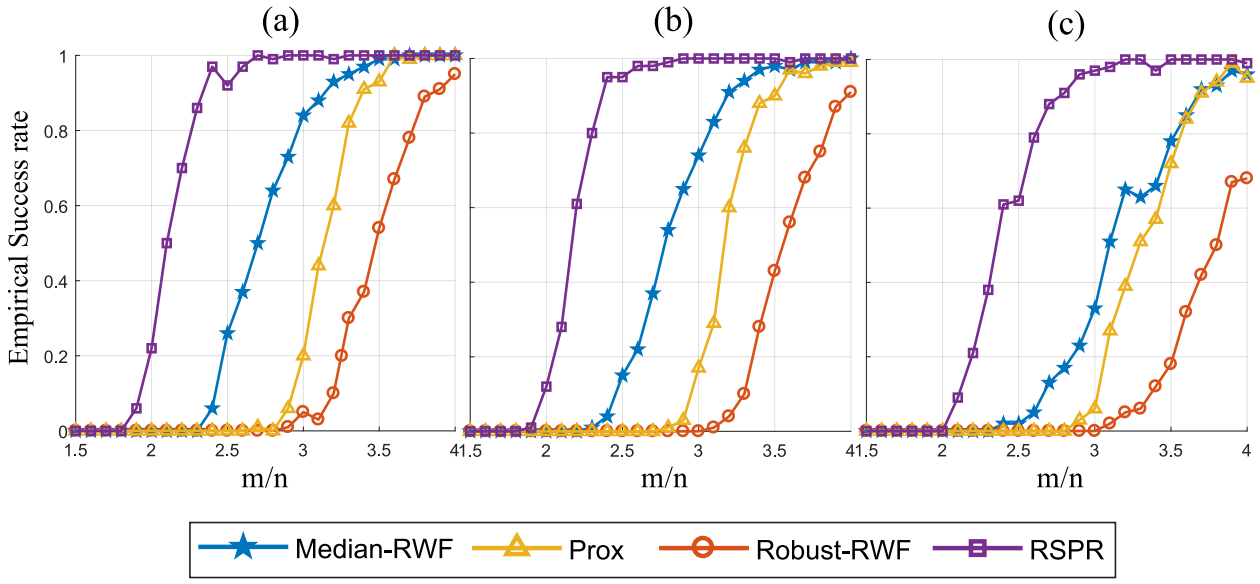
Note: Horizontal axis indexes outliers fraction s while vertical axis indexes the measurement ratio m/n.

6.3.3.1. Performance of the Initialization strategies. The first experiment shows the performance of OPI, used in the proposed method, compared with initializations proposed to deal with sparse outliers such as the Median-RWF Zhang et al. (2018), Robust-RWF Chen et al. (2017) and Small initialization Duchi and Ruan (2017). The number of measurements was fixed as $m = 3n$ and each measurement y_k can be independently corrupted with probability $\alpha \in [0,0{,}4]$. For the first outliers case, set $y_k = 0$, which is more difficult for OPI. In case 2, the outliers $o_k \sim \mathscr{U}(0, o_{max})$ were randomly generated from a uniform distribution. Figure 26 shows the results for different cases. Specifically Figure 26 (a-c) show the case 2 for $o_{max} = 0{,}5\|\mathbf{x}\|_2$, $o_{max} = \|\mathbf{x}\|_2$ and $o_{max} = \|\mathbf{x}\|_\infty \|\mathbf{x}\|_2$, respectively, and Figure 26 (d) shows the case 1. It can be observed that for these scenarios OPI outperforms the other initializations.

Case 1 . In this experiment, the case when the measurements vector and the outliers are independent was evaluated. For this, each measurement $y_k = 0$ can be corrupted with probability of $\alpha \in [0,0{,}3]$ independently. Figure 27 summarizes the success rate of the different methods varying the number of measurements and the outliers fraction. White squares indicate that $100\,\%$ of times, the methods solved the problem satisfactorily, while the black squares indicate $0\,\%$ of success rate. Figure 27 shows two points of reference, where the proposed method obtains betters results compared with the other methods. Specifically, when the number of measurements and the number of outliers increases the proposed method is more robust.

Case 2 . Numerical tests are conducted to demonstrate the robustness of the proposed method under the second outliers scenario. The outliers $o_k \sim \mathscr{U}(0, o_{max})$ were randomly gene-

Figure 28. Empirical success rate versus number fo measurements with $n = 200$ *and outliers fraction* $\alpha = 0{,}1$

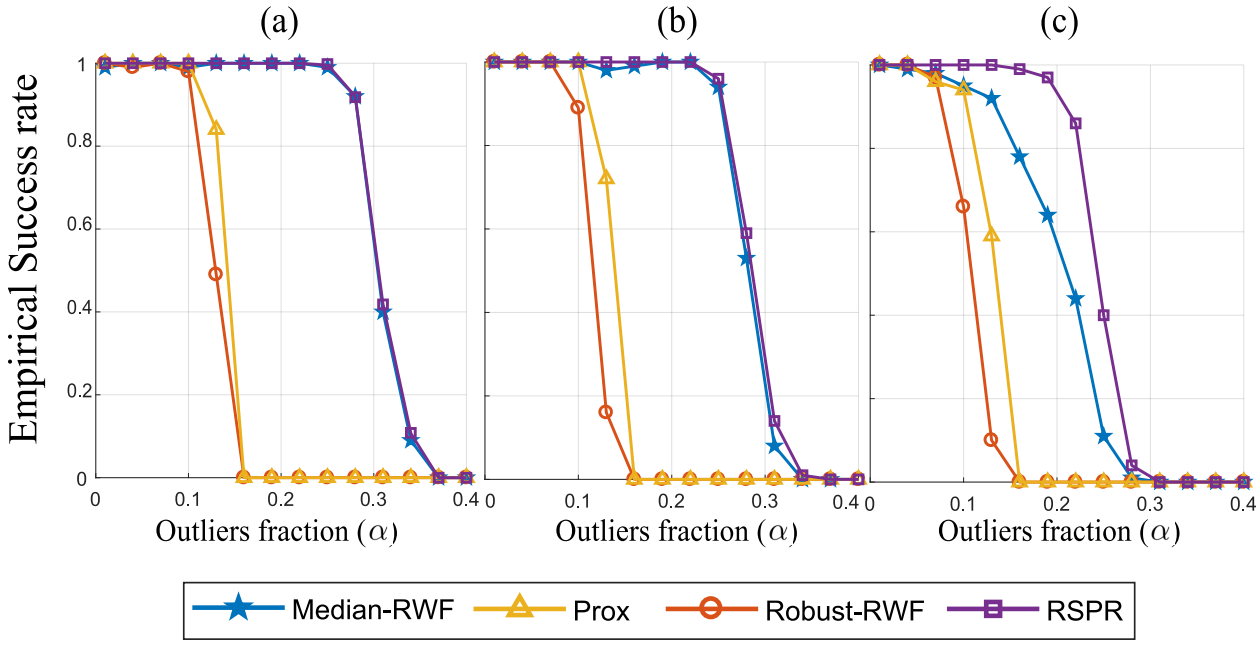


Note: for (a) $o_{max} = 0{,}1\|\mathbf{x}\|_2$, (b) $o_{max} = \|\mathbf{x}\|_2$ and (c) $o_{max} = \|\mathbf{x}\|_\infty \|\mathbf{x}\|_2$.

rated from a uniform distribution and the measurements were generated using the model in (90). Three maximum amplitudes of the outliers were evaluated in this test $o_{max} = 0{,}1\|\mathbf{x}\|_2$, $o_{max} = \|\mathbf{x}\|_2$ and $o_{max} = \|\mathbf{x}\|_\infty \|\mathbf{x}\|_2$. Figure 28 shows the sampling complexity of the phase retrieval method with $\alpha = 0{,}1m$ of outliers. Figure 28 suggests that the proposed method RSPR requires less number of measurements to recover the signal $\mathbf{x}$ in comparison with its robust competitive algorithms. Specifically, it can be observed in fig. 28 (b) that the proposed method needs 25 %, 29 % and 38 % less measurements compared with Median-RWF, Robust-RWF and Prox, respectively.

Additionally, in order to evaluate the robustness of the proposed method, set $m/n = 4$, and the outliers fraction is varied from 0.01 to 0.4 with a step size of 0,03. Results are summarized in Fig.29, where it can be observed that the proposed method has similar performance to Median-

Figure 29. Empirical success rate versus outliers fraction (α), with $n = 200$ and $m = 4n$



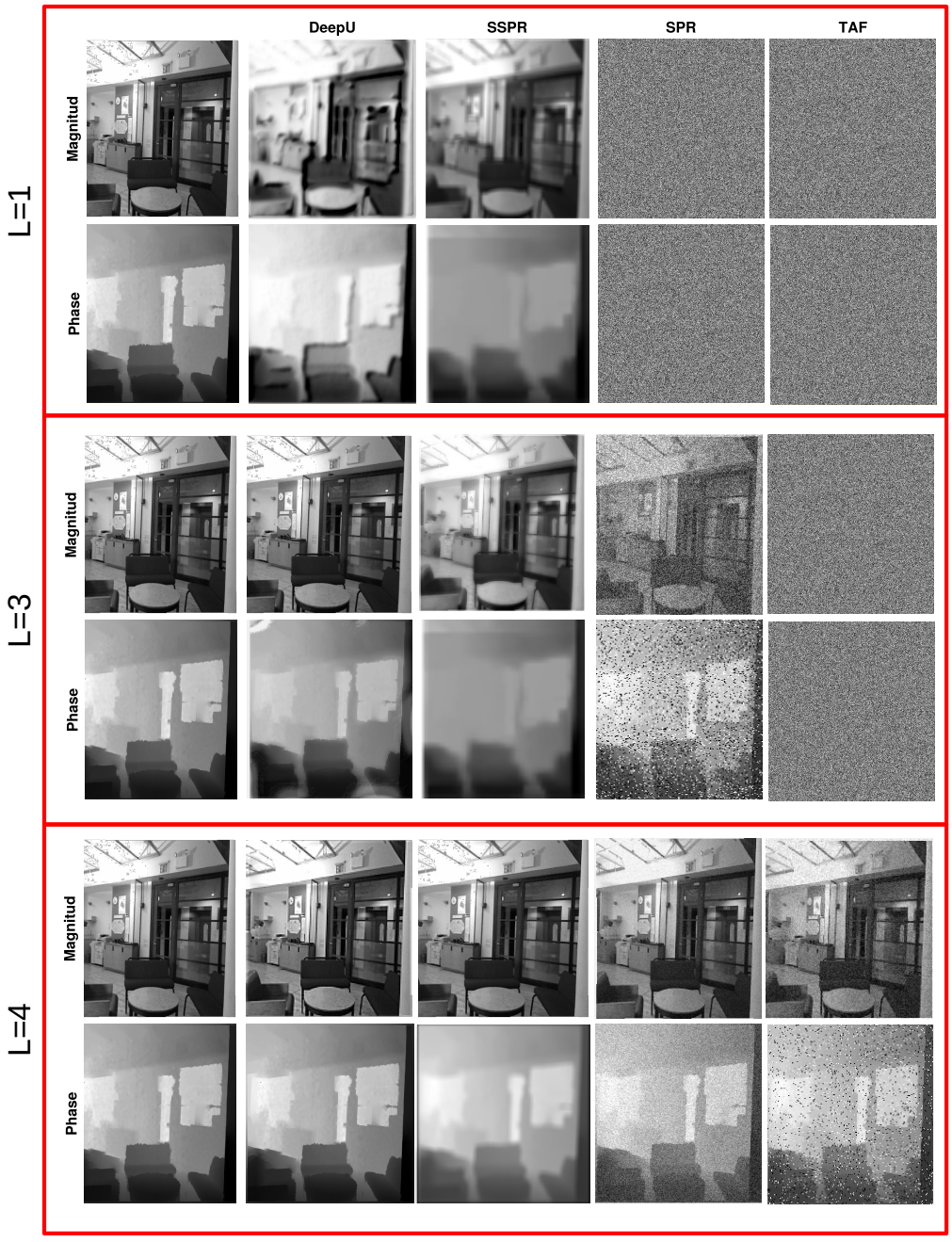
Note: for (a) $o_{max} = 0{,}1\|\mathbf{x}\|_2$, (b) $o_{max} = \|\mathbf{x}\|_2$ and (c) $o_{max} = \|\mathbf{x}\|_\infty \|\mathbf{x}\|_2$.

RWF for small values of o_{max}. However, when the value of the outliers magnitude increases, the proposed method outperforms the other phase retrieval algorithms. Specifically, from fig.29 (c), it can be observed that for $\alpha = 0{,}16$ the success recovery of the proposed method is 20 %, 98 % and 97 % more probable compared with Median-RWF, Robust-RWF and Prox, respectively. Thus, these experiments show a better statistical performance of the RSPR under sparse outliers.

6.4. Deep Unrolled Recovery Network

The deep unrolled recovery network is illustrated in Fig.17. Particularly, this thesis used the E2E formulation for coupled design the CA. Similar to the previous section, the NYU Depth Dataset Silberman et al. (2012) is employed for evaluating the proposed deep approach, which contains 1449 RGB images with depth maps of 15 discretization levels; here, 80 %, 10 %, and

Figure 30. Visual representation of the proposed Deep Unrolling method compared with the state-of-the-art method



Note: for a different number of snapshots.

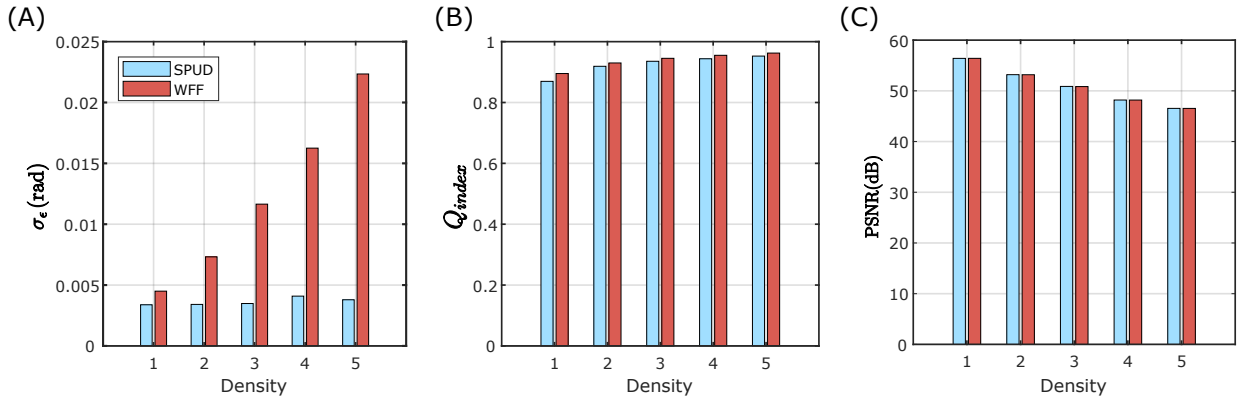
10 % of images were selected for training, testing, and validating, respectively. All images were resized to 256×256 pixels; then, the RGB images were converted to a grayscale version and normalized to simulate the amplitude information. On the other hand, the depth maps were scaled in the range $[-\pi, \pi]$ to simulated the phase information. This thesis evaluated the designed, coded apertures, using random variables $d_1 \in \{0, 1\}$ and this thesis compared with the results of SPR, TAF, and TWF. The Unet-based DNN Bacca et al. (2021) with skip connections at every pooling operation is implemented as a prior network for the Magnitude and Phase, respectively, i.e., two Unet was used. The MSE is used as the main loss function. These experiments were tested using $Ł = 1, 3, 4$ number of snapshots for the proposed method with 30dB of SNR. Figure 30. shows the visual results of a testing image. There it can be observed that the proposed method obtain good results even at one snapshot.

6.5. SPUD

The performance of the proposed phase unwrapping method was evaluated with numerically simulated data and compared to a denoising plus phase unwrapping strategy. The denoising stage is performed using the 2D Windowed Fourier Transform filter (WFF) Kemao (2004, 2007); Kemao et al. (2008), which was shown to outperform the state-of-the-art denoising algorithms in terms of phase error Montresor and Picart (2016). For the phase unwrapping stage, this thesis used the least-squares DCT closed solution. This method is referred to from now on as WFF+LSPU. Suggested parameter settings for WFF were provided in Kemao et al. (2008).

The motivation for these experiments is to find out if, for mild phase noise, the proposed method can avoid the use of costly procedures such as WFF. For this reason, this thesis does

Figure 31. Average phase restoration performance results of SPUD and WFF+LSPU. SPUD outperforms WFF+LSPU



Note: in terms of (A) restored phase error (σ_ε), and performs at par with WFF+LSPU in terms of (B) Quality index and (C) PSNR.

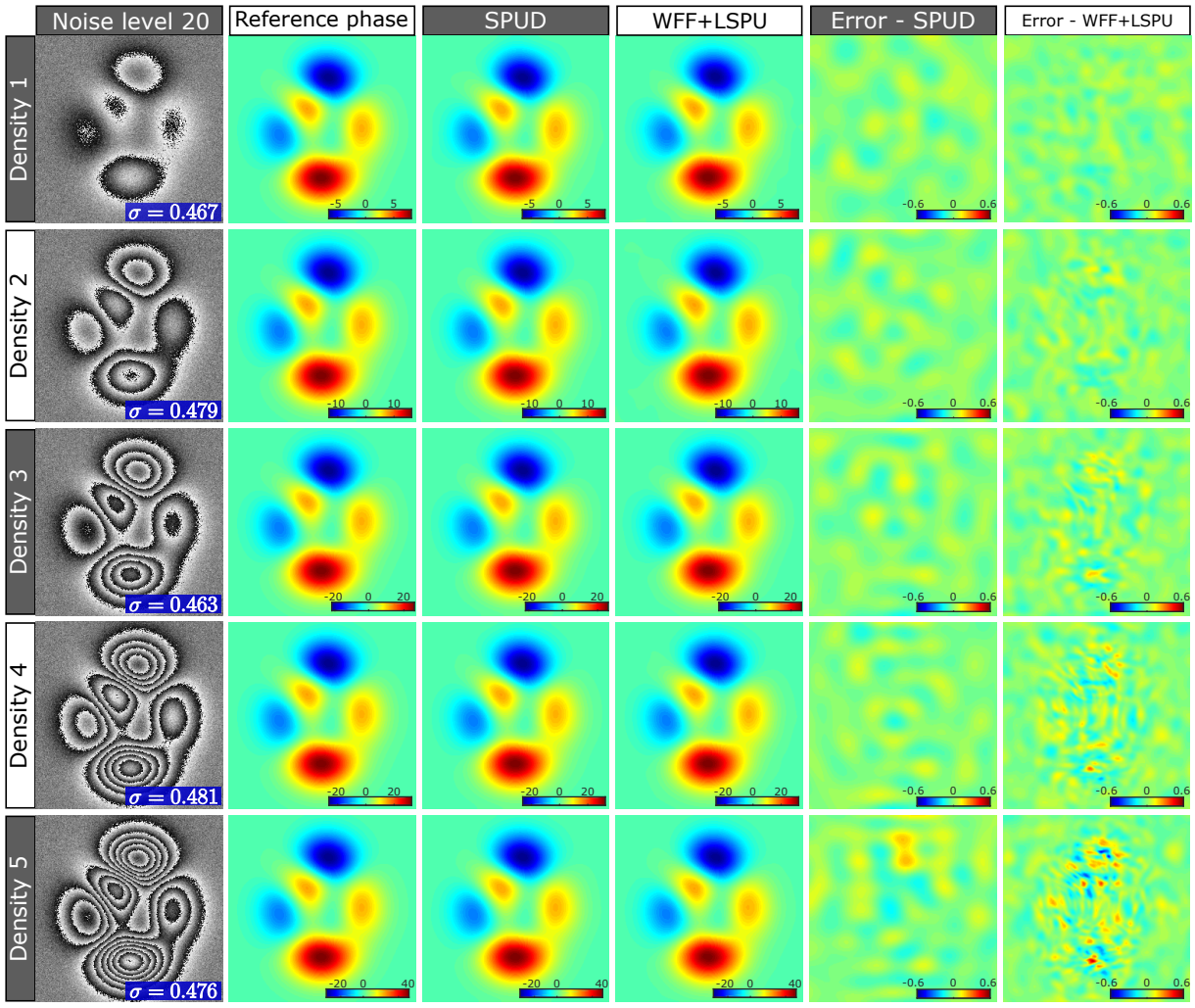
not compare against other noise-robust phase unwrapping methods, like PUMA Bioucas-Dias and Valadao (2007), that have been shown to require additional denoising stages Hongxing and Lingda (2014).

Fig. 31 summarizes the performance of the SPUD algorithm and WFF+ LSPU for the entire dataset and the three metrics explained above. This thesis plots the average values of σ_ε, Q_{index}, and PSNR for the 20 simulated noise levels at each fringe density. From Fig. 31 (A), the proposed method showed the best performance in terms of phase errors compared to WFF+LSPU regardless of fringe density. Whereas from Figures 31 (B) and (C), the SPUD method performs at par with respect to WFF+LSPU in terms of Q_{index} and PSNR.

To visualize the performance results, in Fig. 32, this thesis summarizes the outputs from the SPUD algorithm and WFF+LSPU for the different five phase densities and the highest noise level.

Both methods effectively reduced the noise and correctly unwrapped the phase. However, the phase errors obtained by the proposed method are lower than the errors obtained from WFF+LSPU. Note that the phase errors from WFF+LSPU tend to increase with higher fringe densities at the

Figure 32. 2D representation of the SPUD and WFF+LSPU performance for the five phase densities and the noise level 20.



Note: The SPUD error maps are smoother and the errors are randomly distributed, whereas the WFF+LSPU error maps concentrate high errors in the vicinity of high density fringes.

same noise level, whereas the phase errors from SPUD remain approximately constant regardless of fringe density. Additionally, the errors from SPUD are randomly distributed throughout the phase map, which is a desired property of denoising methods Buades et al. (2011), while the errors from WFF+LSPU are concentrated near high fringe density regions. The quantitative results from Fig. 32 are shown in Table 3 for the three performance metrics. As in the average performance scores, SPUD outperforms WFF+LSPU in terms of phase error σ_ε, and has a comparable performance with respect to WFF+LSPU in terms of Q_{index} and PSNR. The high Q_{index} and PSNR values show the restoration capabilities of both methods.

Table 3. Quantitative assessment of the phase estimation quality in Fig. 32. In bold typeface the values where SPUD performance is superior.

		SPUD			WFF + LSPU		
Density	σ	σ_ε (rad)	Q_{index}	PSNR (dB)	σ_ε (rad)	Q_{index}	PSNR (dB)
1	0.467	0.0059	0.838	56.33	0.0057	0.849	56.36
2	0.479	**0.0067**	0.878	**53.21**	0.0092	0.899	53.19
3	0.463	**0.0069**	0.903	**50.95**	0.0129	0.916	50.83
4	0.481	**0.0058**	**0.945**	**48.21**	0.0187	0.924	48.19
5	0.476	**0.0161**	0.925	46.47	0.0266	0.934	46.54

6.5.1. ***Execution time assessment.*** From the above experiments, it can be concluded that the SPUD method obtains a comparable result with WFF+LSPU. However, the main advantage of the proposed method is its low computational complexity. To illustrate that, this thesis evaluates the execution time of SPUD and WFF+LSPU on a personal computer (PC) with Windows 7 (2.4 GHz i7 intel processor, 8 GB RAM) and MATLAB R2017a. In this experiment, this

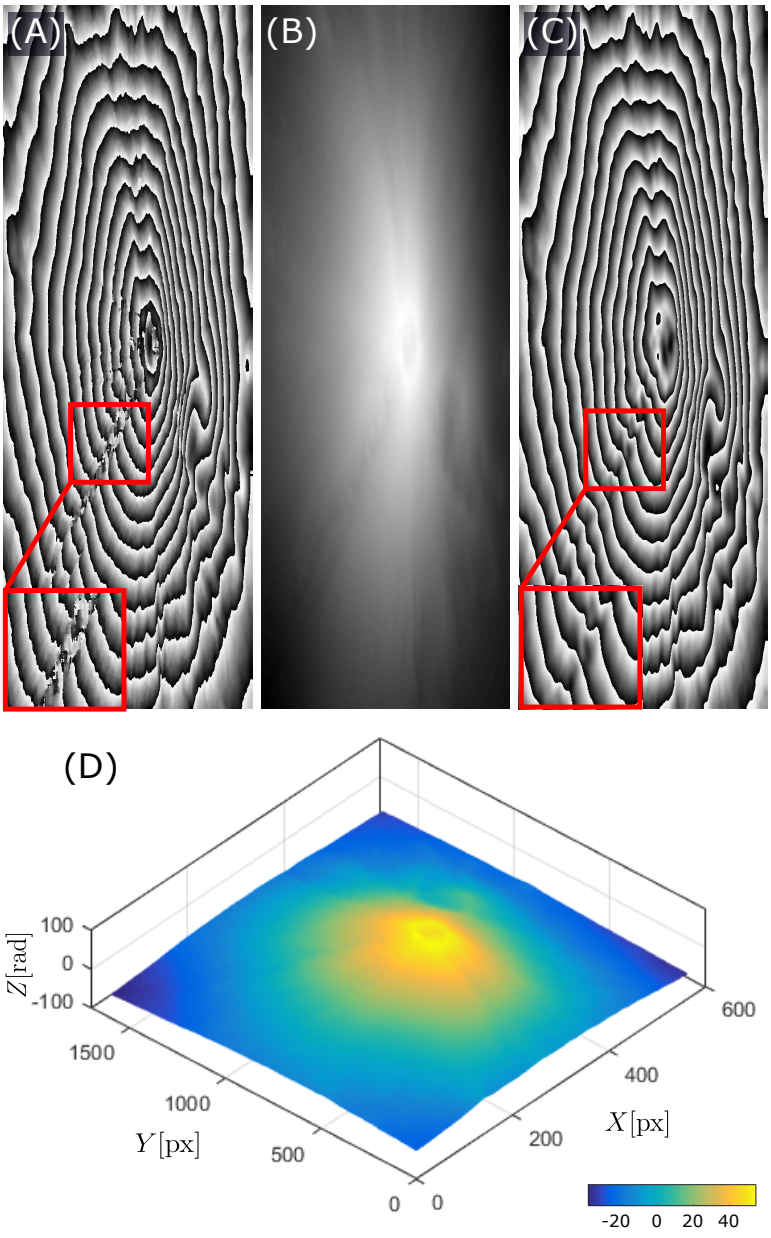
Table 4. Execution time comparison for different array sizes. Time measurements in seconds.

Array Size (pixels)	SPUD (double precision)	WFF+LSPU (double precision)	WFF (GTX295 GPU, single precision)
256× 256	0.0069	24.7215	0.25
512× 512	0.0857	104.2615	0.93
1024× 1024	0.2326	755.8024	3.60

thesis uses three-phase maps with the same phase density and noise level but with different sizes (number of pixels). Table 4 shows the execution time results for both methods in the phased restoration of array sizes: 256×256, 512×512, and 1024×1024. It can be noticed that WFF+LSPU is several orders of magnitude slower than SPUD, and this is mainly due to the high complexity in the denoising stage with WFF. However, since there are GPU implementations of WFF, execution times for the processing of fringe patterns of the same size as reported in Gao et al. (2009) are included. Even in this scenario, the non-optimized MATLAB implementation of the proposed method in double precision (that includes denoising and unwrapping) is one to two orders of magnitude faster than the GPU implementation of WFF in single precision (only the denoising stage without unwrapping).

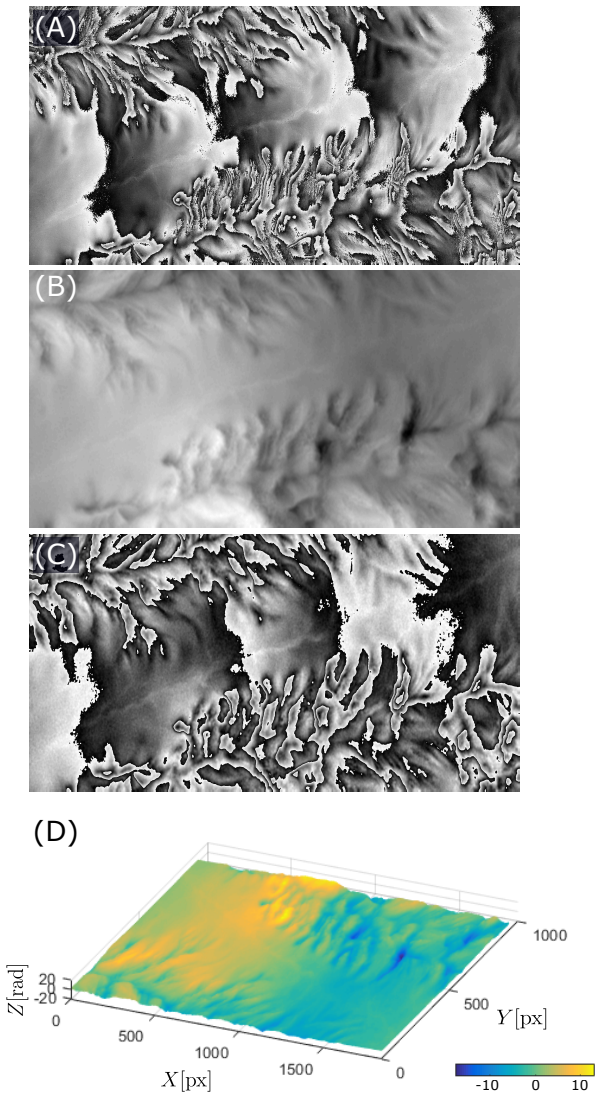
6.5.2. ***Experimental Results*** **.** The SPUD algorithm was evaluated on two wrapped phases from InSAR provided as open-source data in Refs. Hernandez-Lopez et al. (2018); SAR-EDU (2019). Figure. 33(A) shows the first interferometric phase of size 561×1591 pixels, which corresponds to the shape of a volcano. Note that the phase map contains regions of phase dislocations. In this situation, filtering by WFF has proven to be unhelpful Meng et al. (2012). Following the threshold formula, we set $\lambda = 0{,}5\sqrt{2\log(MN)}$ for these experiments.Figure. 33(B)

Figure 33. Phase unwrapping by SPUD from an interferometric wrapped phase of size **1591** $\times$ **561** *pixels.*



Note: (A) Wrapped phase. (B) Unwrapped phase. (C) Re-wrapped values of the unwrapped phase compared with (A). (D) Mesh of the unwrapped phase. The red box indicates a region of phase dislocations, with a zoomed view.

Figure 34. Phase unwrapping by SPUD from an interferometric wrapped phase of size **1065** $\times$ **2032** *pixels.*



Note: (A) Wrapped phase. (B) Unwrapped phase. (C) Re-wrapped values of the unwrapped phase compared with (A). (D) Mesh of the unwrapped phase.

shows the unwrapped and restored phase map obtained by SPUD. Since the dynamic range of the unwrapped result is large, we re-wrap the unwrapped values for visual comparison (Fig. 33(C)), as suggested in Ref. Ghiglia and Pritt (1998). The proposed algorithm removes the regions of pha-

se dislocations from the restored phase map by the smoothing constraint without increasing the computational complexity. Additionally, the unwrapped solution seems congruent with the original data, and no propagation errors are evident. Figure. 33(D) shows the mesh of the unwrapped phase. The processing time in this experiment was 0.504 s.

Figure. 34(A), shows the second wrapped phase map of size 1065×2032 pixels, which corresponds to a region over Phoenix, Arizona, USA, scanned by the Canadian satellite system, RADARSAT-2. The phase map describes a complex topographic area and exhibits noise. Figure. 34(B) shows the unwrapped and restored phase with the proposed method. The re-wrapped phase map is shown in Fig. 34(C). Observe that the effect of noise was minimized with the structural information substantially preserved. The mesh of the unwrapped phase is shown in Fig. 34(D). The processing time for this experiment was 1.04 s.

7. Extension to Compressive Spectral Imaging

This chapter presents some resulting work on compressive spectral imaging (CSI) using the mathematical concepts addressed in this thesis, mainly in the recovery methods and the coded aperture design.

The CSI sensing paradigm acquires 2D multiplexed projections of a three-dimension scene instead of directly acquire all voxels, resulting in image compression via hardware. The spatial-spectral data cube is represented as $\mathscr{F} \in \mathbb{R}^{M\times N\times L}$ with $M \times N$ spatial dimensions, L spectral bands, and $\mathbf{f} \in \mathbb{R}^{MNL}$ denotes the vector representation of the spectral image. Thus, the system matrix model can be expressed as

$$y = \mathbf{Hf}, \tag{137}$$

where $\mathbf{y} \in \mathbb{R}^m$ represent the measurements with $m << MNL$ and $\mathbf{H} \in \mathbb{R}^{m\times MNL}$ represents the linear sensing matrix. Even though CSI yields efficient sensing, a reconstruction process from the compressed measurements is needed since it finds a solution to an under-determined system.

7.1. Compressive Reconstruction:

7.1.1. ***Compressive Spectral Image Reconstruction using Deep Prior and Low–Rank Tensor Representation Bacca et al. (2021).*** The goal in CSI is to recover the spectral image $\mathscr{F} \in \mathbb{R}^{M\times N\times L}$ from the compressive measurements y. A tensor formulation for addressing

this problem is described below

$$\begin{aligned} \underset{\mathscr{Z}_o}{\text{minimize}} \quad & \frac{1}{2}\|y - \mathbf{H}\text{vect}(\mathscr{F})\|_2^2 + \lambda\phi({\mathscr{Z}'}_o) \\ \text{subject to} \quad & \mathscr{F} = {\mathscr{Z}'}_o \times_1 \mathbf{U}' \times_2 \mathbf{V}' \times_3 \mathbf{W}', \end{aligned} \tag{138}$$

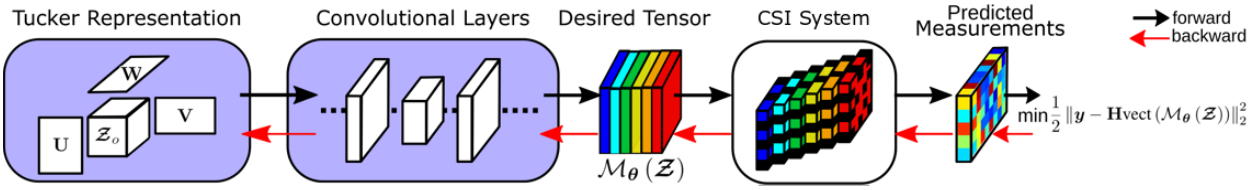
where the matrices $\mathbf{U}' \in \mathbb{R}^{M\times M}$, $\mathbf{V}' \in \mathbb{R}^{N\times N}$ and $\mathbf{W}' \in \mathbb{R}^{L\times L}$ are fixed and known orthogonal matrices, which usually are the matrix representation of the Wavelet and the Discrete Cosine transforms; ${\mathscr{Z}'}_o$ is the representation of the spectral image in the given basis and $\phi(\cdot) : \mathbb{R}^{M\times N\times L} \to \mathbb{R}$ is a regularization function that imposes particular image priors with λ as the regularization parameter Figueiredo et al. (2007).

Unlike the hand-craft priors as sparsity Arce et al. (2014), this thesis explores the power of some deep neural networks as image generators that map a low-dimensional feature tensor $\mathscr{Z} \in \mathbb{R}^{M\times N\times L}$ to the image as

$$\mathscr{F} = \mathscr{M}_\theta(\mathscr{Z}), \tag{139}$$

where $\mathscr{M}_\theta(\cdot)$ represents a deep network, with θ as the net-parameters. To ensure a low-dimensional structure over the feature tensor, this thesis used the Tucker representation, i.e., $\mathscr{Z} = \mathscr{Z}_o \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W}$ with $\mathscr{Z}_o \in \mathbb{R}^{M_\rho\times N_\rho\times L_\rho}$ as a 3D low dimensional tensor, with $M_\rho < M$, $N_\rho < N$ and $L_\rho < L$. This representation, in the input of the network, aims to maintain the 3D structure of the spectral images, exploits the inherent low-rank of this data Wang et al. (2017d); León-López and Fuentes (2020), and also implicitly constraint the output $\mathscr{F}$ in a low dimensional manifold via the archi-

Figure 35. Visual representation of the proposed deep neural scheme, where the boxes with background color represent the learning parameters.



Note: The white box stand for the non-trainable CSI system, and the non-box blocks represent the outputs of the layers.

tecture and the weights of the net Wu et al. (2019). It is worth highlighting that, unlike Wang et al. (2017d); León-López and Fuentes (2020), this thesis does not satisfy the low-rank structure in the recovered spectral image (output of the network). Instead, this thesis imposes Tucker decomposition on the input network, which expects that after some convolution layer, extract some non-linearity features present in the SI.

This thesis is focused on a blind representation, where instead of having a pre-training network or massive amount of data to train this deep neural representation, this thesis expresses an optimization problem which learns the weight θ in the generative network $\mathscr{M}_\theta$ and also the tensor feature $\mathscr{Z}$ with its Tucker representation elements as $\mathscr{Z}_o, \mathbf{U}, \mathbf{V}$ and $\mathbf{W}$. All the parameters of this optimization problem are randomly initialized, and the only available information is the compressive measurements and the sensing model, i.e., the optimization problem is data training independent. In particular, this thesis explores the prior implicitly captured by choice of the generator network structure, which is usually composed of convolution operations, and the importance of the low-rank representation feature; therefore, the proposed method consists of solving the following

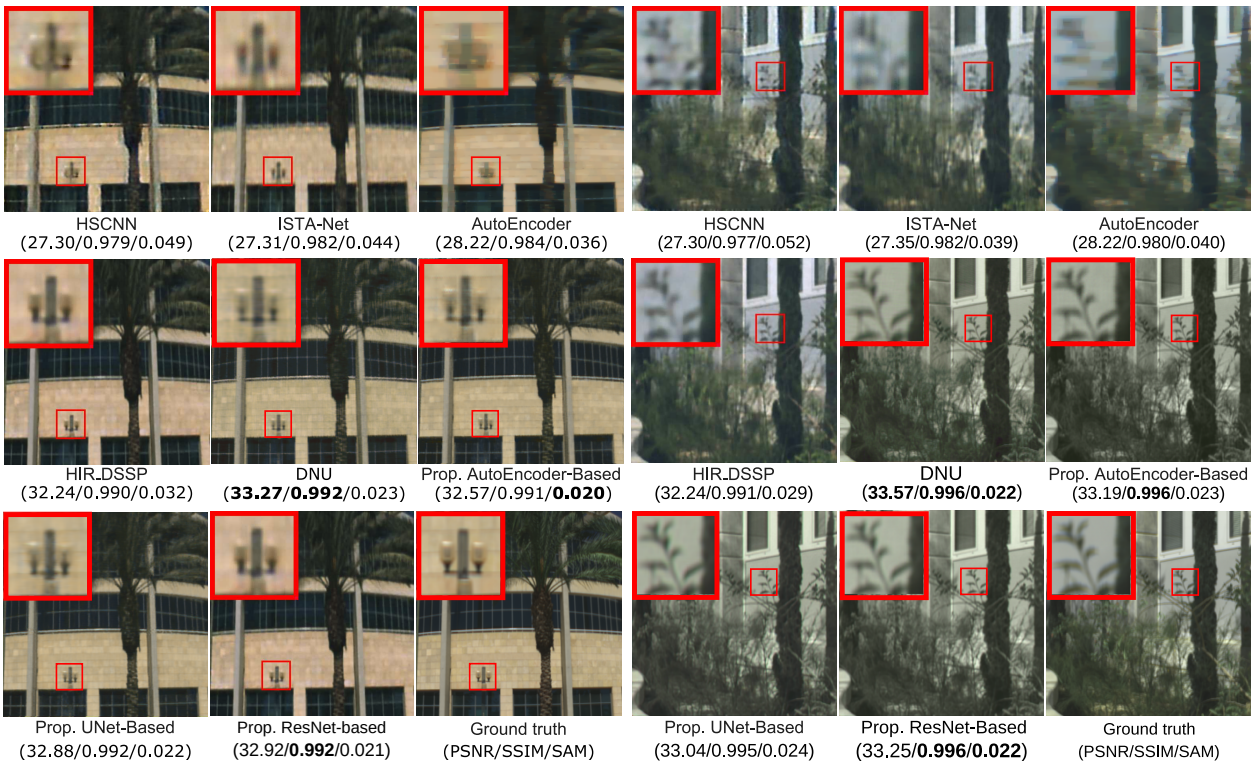
optimization problem

$$\begin{aligned} \underset{\theta,\mathscr{Z}_o,\mathbf{U},\mathbf{V},\mathbf{W}}{\text{minimize}} \quad & \frac{1}{2}\|y-\mathbf{H}\text{vect}\left(\mathscr{M}_\theta\left(\mathscr{Z}\right)\right)\|_2^2 \\ \text{subject to} \quad & \mathscr{Z}=\mathscr{Z}_o\times_1\mathbf{U}\times_2\mathbf{V}\times_3\mathbf{W}, \end{aligned} \tag{140}$$

where the recovery is $\mathscr{F}^*=\mathscr{M}_{\theta^*}(\mathscr{Z}_o^*\times_1\mathbf{U}^*\times_2\mathbf{V}^*\times_3\mathbf{W}^*)$. This optimization problem can be solved using an end-to-end neural network framework, as shown in Fig. 35. In this way, the input, that is common in all neural networks, is replaced with a custom layer with $\mathscr{Z}_o,\mathbf{U},\mathbf{V},\mathbf{W}$ as learnable parameters, which construct the low-rank Tucker representation of $\mathscr{Z}$, then this tensor $\mathscr{Z}$ is refined with convolutional layers via $\mathscr{M}_\theta(\mathscr{Z})$; these optimization variables are represented by the first two blue-blocks in the Fig. 35. The final layer in the proposed method is a non-training layer which models the forward sensing operator $\mathbf{H}\text{vect}\left(\mathscr{M}_\theta\left(\mathscr{Z}\right)\right)$ to obtain the compressive measurements y as the output of the net. Therefore, the problem in (140) can be solved with state-of-the-art deep learning optimization algorithm, such as, stochastic gradient descent. Once the parameters are optimized, the desired SI is recovered just before the non-trainable layer labeled as ÇSI systemïn Fig. 35.

7.1.1.1. Simulations and Results. Although the proposed method does not need data to work, this test compares its results with the deep learning approaches to demonstrate the quality achieved. In particular, this thesis uses five learning-based methods for comparison: HS-CNN Xiong et al. (2017), ISTA-Net Zhang and Ghanem (2018), Autoencoder Choi et al. (2017); HIR-DSSP Wang et al. (2019) and DNU Wang et al. (2020). These methods were trained using the

Figure 36. Two reconstructed scenes using the 5 learning-based methods.



Note: Three are the variations of the proposed method, i.e., (AutoEncoder, UNet, and ResNet)-Based.

public ICVL Arad and Ben-Shahar (2016), Harvard Chakrabarti and Zickler (2011), and KAIST Choi et al. (2017) hyperspectral image data-sets using their available codes and following the principles in Wang et al. (2018d, 2019) to partition the training and testing sets; the sensing process was evaluated for a single snapshot with 30 dB of SNR, according to Wang et al. (2020). For this section, ResNet-based, AutoEnconder-Based, and UNet-based were used as the Convolutional layer in the proposed method with $\rho = \{0{,}5, 0{,}7, 0{,}7\}$, respectively. Two testing images of 512×512 of spatial resolution and 31 spectral bands were chosen to evaluate the different methods, and the reconstruction results and ground truth are shown in Fig. 36. It can be observed that the two variants

Table 5. Computational complexity of the deep learning and the proposed methods measured as mean time in seconds of 5 trials.

Methods	HSCNN	ISTA-Net	AutoEncoder	HIR-DSSP	DNU	Prop. AutoEncoder	Prop. UNet	Prop. ResNet
GPU Time [s]	8.708	3.224	575.421	8.397	**2.744**	137.375	278.0411	135.834
CPU Time [s]	72.174	27.154	3948.421	68.214	**20.727**	1084.154	2224.145	997.156

of the proposed method outperform in visual and quantitative results to HSCNN, ISTA-Net, AutoEnconder, HIR-DSSP, up to $(5/0{,}030/0{,}020)$ in terms of (PSNR/SSIM/SAM), respectively, and show comparable/close results with respect to the DNU method, which is the best deep learning method. To make a fair run-time comparison of the different methods, all the recovery approaches were running in an Intel (R) Xeon (R) CPU 2.80 GHz. Additionally, since all deep learning methods are implemented to use GPU, it is also run it Google Colab source using an NVIDIA Tesla P100 PCIe 16 GB. Table 5 shows the running time for reconstructing one spectral image from the compressive measurements. Notice that the proposed methods are iterative; therefore, it employed 2,000 iterations which offers a stable convergence. Although the execution time to obtain a spectral image is longer than most deep learning methods, the proposed methods have the advantage that it does not require training, i.e., only the compressive measurements are available for the proposed approach.

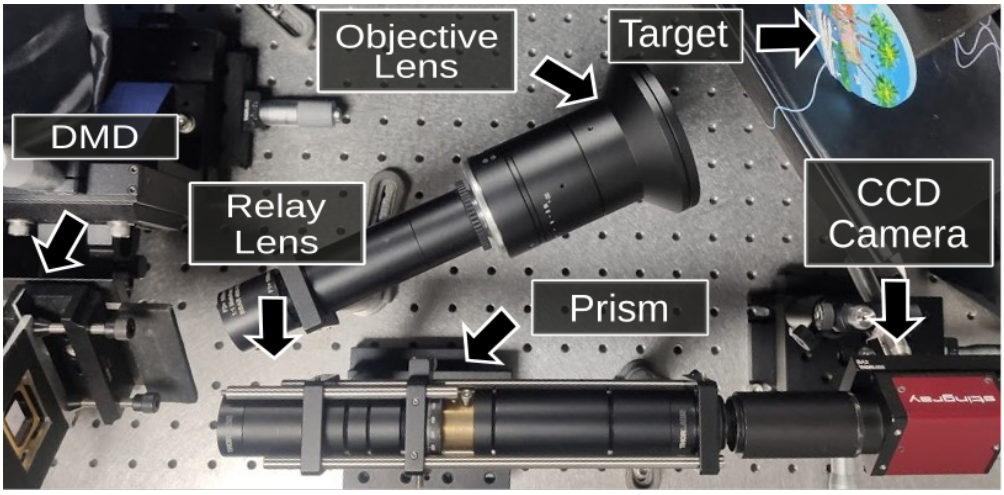
7.1.1.2. Validation in a Real Testbed Implementation. This section evaluates the proposed method with real measurements acquired using a testbed implementation. For this section, the ResNet-based model was used with ($\rho = 0{,}4$), and learning rate $1e-3$. Specifically, two different scenarios of compressed projections were assessed, which are described as follows.

This scenario was carried out for one snapshot of the CASSI testbed laboratory implementation depicted in Fig. 37. This setup contains a 100-*nm* objective lens, a high-speed digital

micro-mirror device (DMD) (Texas Instruments-DLI4130), with a pixel size of $13,6\mu m$, where the CA is implemented, an Amici Prism (Shanghai Optics), and a CCD (AVT Stingray F-145B) camera with spatial resolution 1388×1038, and pitch size of $6{,}45\mu m$. The CA, spatial distribution for the snapshot, comes from blue noise patterns, i.e., this CA is designed according to Correa et al. (2016a). The coding and the scene were implemented to have a spatial resolution of 512×512 pixels and $L = 13$ as the resolvable bands. This thesis decided to compare with PnP-ADMM, and DIP using this real data.

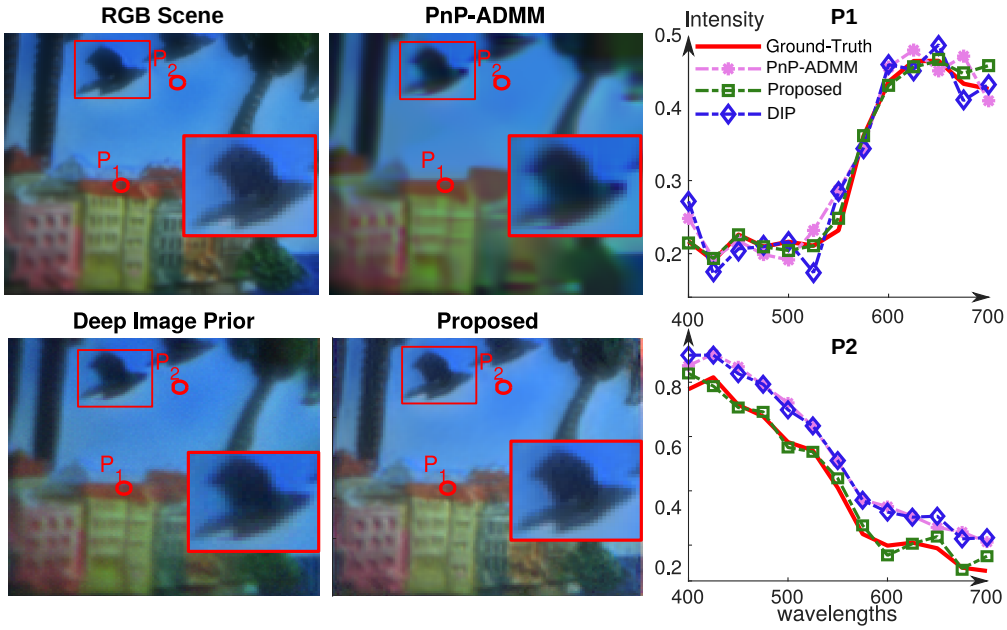
Figure 38 presents the RGB scene obtained with a traditional camera, and the false-colored RGB images corresponding to reconstructed spectral images using the different solvers. Furthermore, the spectral responses of two particular spatial locations in the scene indicated as red points in the images are also included and compared with the spectral behavior using a commercially

Figure 37. Testbed CASSI implementation.



Note: The relay lens focuses the encoded light by the DMD into the sensor after dispersed by the prism.

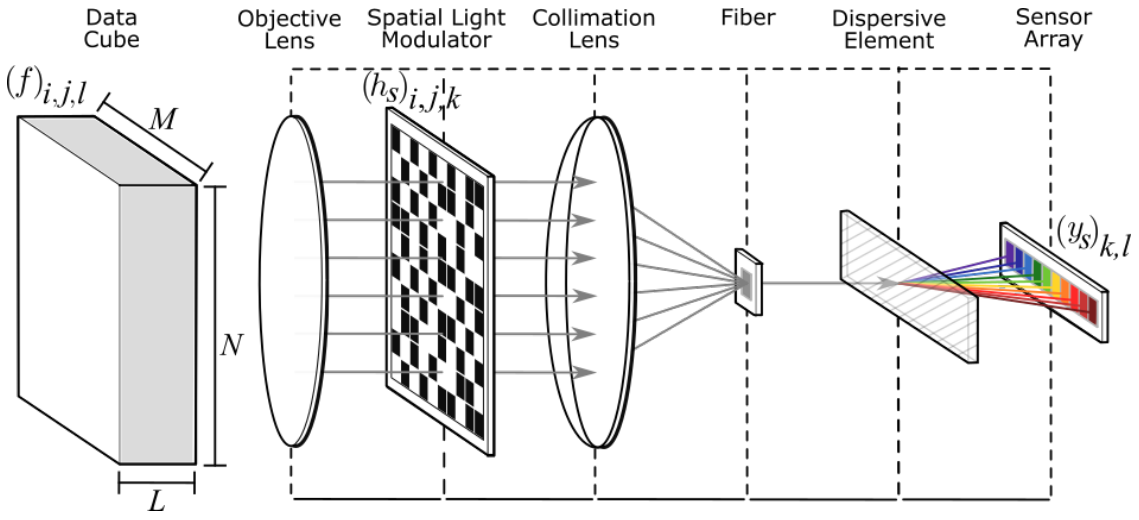
Figure 38. (Left) RGB visual representation of the scene obtained with the different methods, (Right), two spectral signatures of the recovered scenes.



available spectrometer (Ocean Optics USB2000+). The visual results show that the proposed method yield better spatial and spectral reconstruction since the RGB reconstructed is sharper in the proposed scheme, and the spectral signatures are closer to those taken by the spectrometer; this is, the SAM of the normalized signatures obtained from the PnP-ADMM algorithm is 0.188, Deep Image Prior is 0.205, and the SAM associated to the proposed method is 0.120. These numerical results validate the performance of the proposed method with real data for a real CASSI setup using a binary-coded aperture.

7.1.2. ***Non-Iterative Hyperspectral Image Reconstruction from Compressive Fused Measurements Bacca et al. (2019).*** Several camera designs have been proposed for compressive spectral image acquisition Arguello and Arce (2014); Correa et al. (2015). A common point among those is using a random pattern to modulate the spatial information to obtain the measure-

Figure 39. Single Pixel Camera schematic for hyperspectral data acquisition



ment collection. This thesis is focused on the fusion of measurements from the single-pixel camera and the 3D-CASSI scheme.

7.1.2.1. Single Pixel Camera. The single-pixel camera (SPC) for spectral imaging Sun and Kelly (2009); Soldevila et al. (2013) is illustrated in Fig. 39. SPC, in the k-th shot, uses a spatial light modulator which spatially codes all the spectral bands of the data cube using the same block-unblock pattern $(h_s)_{k,i,j}$. Then, the encoded data is projected into a single spatial point where a spectrometer is used as the detector. This system can be modeled as a linear mapping, where all pixels (i, j) of the image $(f)_{i,j,l}$ in each spectral band l are mapped to a single point $(y_s)_{k,l}$. This is expressed mathematically as

$$(y_s)_{k,l} = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (h_s)_{k,i,j} (f)_{i,j,l}, \tag{141}$$

with $(h_s)_{k,i,j} \in \{0,1\}$; where $k = 0,\ldots,K_s-1$ indexes the different captured projections for a total of K_s shots. Equation (141) can be expressed as the linear system

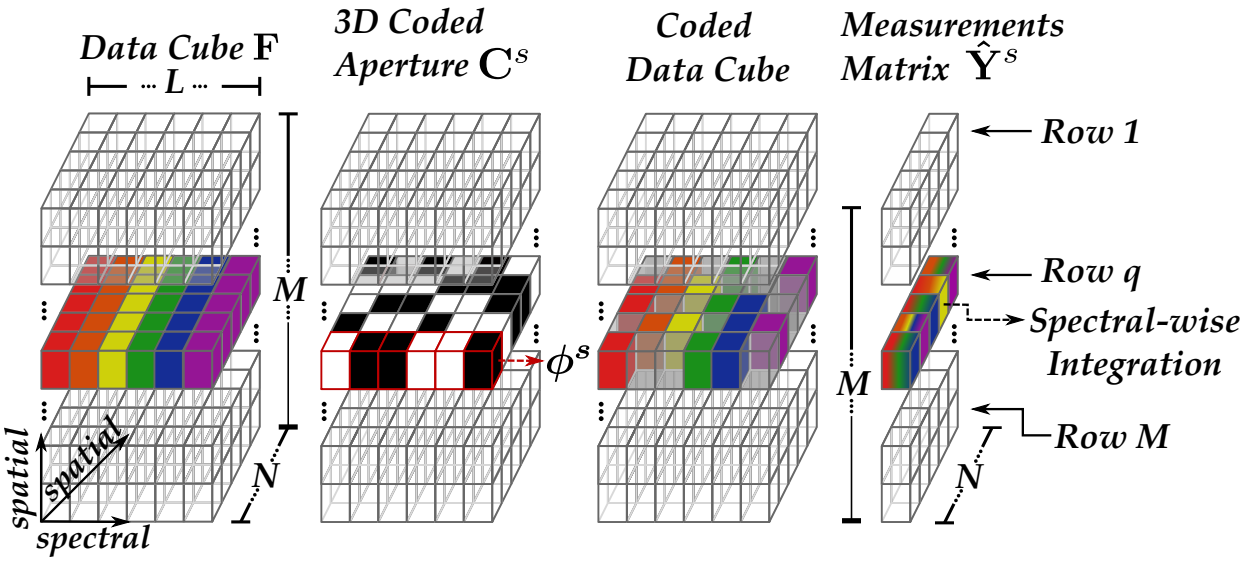
$$\mathbf{Y}_s = \mathbf{H}_s\mathbf{F}, \tag{142}$$

where $\mathbf{H}_s \in \{0,1\}^{K_s\times MN}$ represents the SPC sensing matrix, $\mathbf{Y}_s \in \mathbb{R}^{K_s\times L}$ denotes the observation matrix with K_s shots, and $\mathbf{F} \in \mathbb{R}^{MN\times L}$ is the hyperspectral data cube organized as $\mathbf{F} = [\mathbf{f}_{(1)},\ldots,\mathbf{f}_{(MN)}]^T$, where $\mathbf{f}_{(\ell)} \in \mathbb{R}^{L\times 1}$ represents the spectral signature of the ℓ-th pixel. Notice that the compression rate of the SPC is given by

$$\%_{SPC} = K_s/MN. \tag{143}$$

7.1.2.2. 3D-CASSI Scheme. The spatial-spectral coded compressive spectral imaging system (3D-CASSI) is a CSI sensing scheme that modulates the spectral data cube in the

Figure 40. Schematic of 3D-CASSI sensing approach.



spatial and spectral dimensions simultaneously Cao et al. (2016). Unlike SPC, which encodes all the spectral bands with the same pattern, 3D-CASSI uses a 3D-coded aperture $(h_c)_{k,i,j,l}$ (ensemble of 2D coded apertures), which enables different coding for each spectral band. Specifically, the 3D coded aperture, in the k-th shot, is composed by a set of coding filters $\phi^s \in \mathbb{R}^L$, for $s = 0, \cdots S-1$ where S denote the number of different coding patterns distributed on h_c in a single shot. Note that in general $S < L$Correa et al. (2015). Then, the coded spectral data cube is integrated along the spectral dimension such that each spatial position of the acquired measurements contains the compressed information of a single coded spectral signature as shown in Fig. 40. Works in Gehm et al. (2007); Lin et al. (2014a,b) describe different implementations of the 3D-CASSI scheme (see Cao et al. (2016) for more details).

Mathematically, the output of the sensing process, at the (i, j)-th detector pixel and a specific snapshot k, can be expressed as

$$(y_c)_{k,i,j} = \sum_{l=0}^{L-1} (h_c)_{k,i,j,l} (f)_{i,j,l}. \tag{144}$$

Assuming that K_c shots are taken, the set of compressed measurements from (144) can be arranged in a $K_c \times MN$ matrix

$$\hat{\mathbf{Y}} = \big[[(y_c)_{(0,0,0)}, (y_c)_{(0,1,0)}, \cdots, (y_c)_{(0,0,1)}, \cdots, (y_c)_{(0,M-1,N-1)}]^T,$$

$$\cdots, [(y_c)_{(K_c-1,0,0)}, \cdots, [(y_c)_{(K_c-1,M-1,N-1)}]^T\big]^T$$

, where each column value corresponds to a compressed spectral signature.

On the other hand, if we assume that $K_c = S$, the entries of $\hat{\mathbf{Y}}$ can be rearranged to form a new matrix $\mathbf{Y}_c$, such that each row contains the compressed information acquired with a specific coding pattern ϕ^s Hinojosa et al. (2018). Formally, the rearrangement can be expressed as

$$Y_{s,j} = \hat{Y}_{s',j} \quad \text{if } \hat{Y}_{s',j} = (\phi^s)^T \mathbf{f}_j \quad \forall s', \tag{145}$$

for $s, s' = 0, \cdots, K_c - 1$, where $\mathbf{f}_j$ denotes the j-th spectral signature with $j = 0, \cdots, MN - 1$. This rearrangement, depicted in Fig. 41, preserves the structure of the underlying high dimensional data. Note that this rearrangement is possible only when $K_c = S$, since in this case, it can be guaranteed that, at a specific snapshot, one pixel is encoded only once by a different pattern and, at the end of the sensing procedure, all pixels were encoded by the whole set of K_c coding patterns. Alternatively, define the matrix of K_c coding patterns as $\mathbf{H}_c = \left[\phi^0, \phi^1, \cdots \phi^{K_c-1}\right]^T$ then, the problem of acquiring and rearranging the measurements $\mathbf{Y}_c$ can be succinctly expressed as

$$\mathbf{Y}_c = \mathbf{H}_c \mathbf{F}^T, \tag{146}$$

where the compression rate of this CSI sensing scheme is

$$\%_{3D} = K_c / L. \tag{147}$$

From (143) and (147) the total compression rate given by the fusion of both sets of measurements

is expressed as

$$\%_{total} = \%_{SPC} + \%_{3D} = \frac{K_s L + K_c MN}{MNL}. \tag{148}$$

It should be noted that the total compression rate should not exceed 1. Therefore, the individual compression rates $\%_{SPC}$ and $\%_{3D}$ should be at most 0.5.

7.1.2.3. Reconstruction Algorithm. A fundamental assumption of this thesis is that each spectral vector $\mathbf{f}_j$ lives in a $P-$dimensional subspace $\mathscr{S}$, with $P << L$. This is an approach commonly used in many HSI applications Yang et al. (2010); Bacca et al. (2017); Qian et al. (2011); Nascimento and Dias (2005). Let $\mathbf{E} = [\mathbf{e}_1 \cdots \mathbf{e}_P]$ be a matrix holding a basis for the subspace $\mathscr{S}$ on its columns, where $\mathbf{e}_j \in \mathbb{R}^L$. Therefore, each spectral vector $\mathbf{f}_j$ can be represented (non-uniquely[1]) as

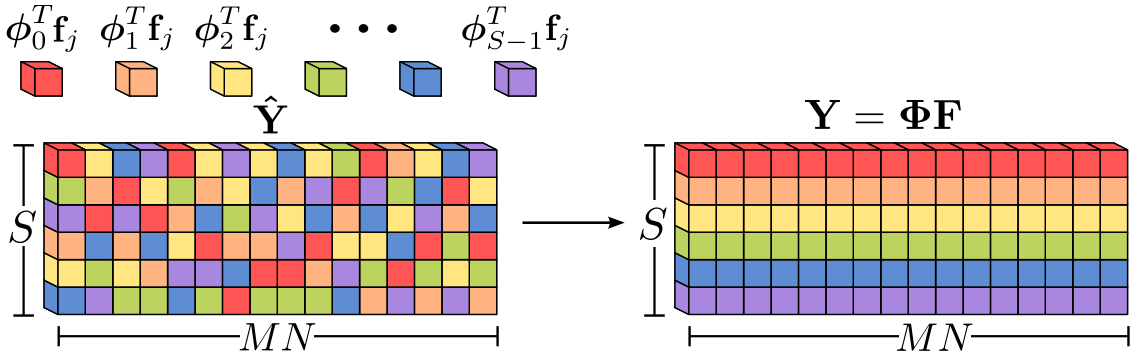
$$\mathbf{f}_j = \mathbf{E}\mathbf{a}_j, \ for\ j = 0, \cdots MN-1, \ \ and \ \ \mathbf{F}^T = \mathbf{E}\mathbf{A}, \tag{149}$$

with $\mathbf{A} = [\mathbf{a}_1 \cdots \mathbf{a}_{MN}]$, where $\mathbf{a}_j \in \mathbb{R}^P$ is the representation coefficient of $\mathbf{f}_j$ with respect to the basis $\mathbf{E}$. This formulation can be applied to the linear mixture model (LMM), since each column of the matrix $\mathbf{E}$ may be interpretable as one endmember, and $\mathbf{A}$ as the corresponding abundance matrix when the sum-to-one constraint (ASC) and the non-negative constraint (ANC) are included in the reconstruction problem Bioucas-Dias et al. (2012). However, this thesis is based on the general case of finding the basis of a low-dimensional subspace and its respective coefficients. Note that, taking (149) into account, (144) can be re-expressed as

---

[1] The matrix $\mathbf{F}^T$ can be rewritten as $\mathbf{F}^T = (\mathbf{E}\mathbf{R})(\mathbf{R}^T\mathbf{A})$ for any rotation matrix $\mathbf{R}$.

Figure 41. Rearrangement of the matrix $\hat{\mathbf{Y}}$ such that the s-th row of $\mathbf{Y}$ contains the compressed measurements acquired with the s-th coding pattern ϕ^s.



$$
\begin{aligned}
(y_c)_{k,i,j} &= \sum_{l=0}^{L-1} (h_c)_{k,i,j,l} \sum_{p=0}^{P-1} (e)_{l,p}(a)_{i,j,p} \\
&= \sum_{p=0}^{P-1} (a)_{i,j,p} \sum_{l=0}^{L-1} (h_c)_{k,i,j,l}(e)_{l,p},
\end{aligned} \tag{150}
$$

which shows that, in the 3D-CASSI scheme, the sensing matrix only affects the basis, while the representation coefficients remain constant. Now, let $\tilde{\mathbf{E}} = \mathbf{H}_c\mathbf{E}$ be the codification basis, then, from (146) and (149) the measurements can be rewritten as

$$\mathbf{Y}_c = \tilde{\mathbf{E}}\mathbf{A}. \tag{151}$$

The matrix $\tilde{\mathbf{E}}$ can be obtained using, for instance, the VCA endmember extraction algorithm Nascimento and Dias (2005) with P endmembers or any other way to find the basis of a subspace such as the SVD decomposition as

$$\Sigma = \mathbf{Y}_c\mathbf{Y}_c^T/MN = \tilde{\mathbf{E}}\mathbf{A}\mathbf{A}^T\tilde{\mathbf{E}}^T = \mathbf{W}\Lambda\mathbf{W}^T, \tag{152}$$

such that, choosing the P eigenvectors corresponding to the P largest eigenvalues as columns of a matrix $\mathbf{W}$, yields $\tilde{\mathbf{E}} = [\mathbf{W}_0, \cdots \mathbf{W}_{P-1}]$. Further, assuming that the columns of $\tilde{\mathbf{E}}$ are linearly independent, the abundance matrix can be exactly recovered as

$$\mathbf{A} = \left(\tilde{\mathbf{E}}^T \tilde{\mathbf{E}}\right)^{-1} \tilde{\mathbf{E}}^T \mathbf{Y}_c. \tag{153}$$

On the other hand, $\mathbf{E}$ cannot be recovered from (151) since it is an undetermined system of equations with $K_c << L$. However, notice that from (141) we have that

$$\begin{aligned}(y_s)_{k,l} &= \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (h_s)_{k,i,j} \sum_{p=0}^{P-1} (e)_{l,p} (a)_{i,j,p} \\ &= \sum_{p=0}^{P-1} (e)_{l,p} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (h_s)_{k,i,j} (a)_{i,j,p},\end{aligned} \tag{154}$$

which shows that this architecture only affects the abundance map. Thus, (154) can be expressed in matrix form as

$$\mathbf{Y}_s = \mathbf{H}_s \mathbf{A}^T \mathbf{E}^T, \tag{155}$$

and taking into account that it can obtain $\mathbf{A}$ from (153), the exact reconstruction of $\mathbf{E}$, provided that the product of $\mathbf{A}\mathbf{H}_s^T$ is full row rank, can be calculated as

$$\mathbf{E} = \mathbf{Y}_s^T \mathbf{H}_s \mathbf{A}^T \left(\mathbf{A}\mathbf{H}_s^T \mathbf{H}_s \mathbf{A}^T\right)^{-1}. \tag{156}$$

Exact reconstruction of $\mathbf{F}$, in the absence of noise, can be achieved under the conditions established in Theorem 6. Algorithm 13 summarizes the steps explained above for the proposed method.

Theorem 6. (**Noiseless exact reconstruction**): Assume that an HSI $\mathbf{F} \in \mathbb{R}^{L\times MN}$, has rank $P \leq$ mín$\{L, MN\}$, such that it can be factorized as $\mathbf{F}^T = \mathbf{EA}$ where $\mathbf{E} \in \mathbb{R}^{L\times P}$ and $\mathbf{A} \in \mathbb{R}^{P\times MN}$. Also, consider the measurement matrices $\mathbf{H}_s \in \{0,1\}^{K_s\times MN}$ and $\mathbf{H}_c \in \{0,1\}^{K_c\times L}$, independently drawn at random from a Bernoulli (1/2) distribution. Then, $\mathbf{F}$ can be exactly recovered from the compressed measurements $\mathbf{Y}_s$ and $\mathbf{Y}_c$ by recovering $\mathbf{A}$ and $\mathbf{E}$ from the solution of (153) and (156), respectively, if $P \leq K_s$ and $P \leq K_c$ shots are taken.

Demostración. The proof of Theorem 6 is in the published article Bacca et al. (2019). □

Algorithm 12 FCSI reconstruction

Input: Measurements $\mathbf{Y}_s, \mathbf{Y}_c$, sensing matrices $\mathbf{H}_s, \mathbf{H}_c$, and

dimension of the subspace P.

1: $\tilde{\mathbf{E}} \leftarrow \text{VCA}(\mathbf{Y}_c)$

2: $\mathbf{A} \leftarrow \left(\tilde{\mathbf{E}}^T\tilde{\mathbf{E}}\right)^{-1}\tilde{\mathbf{E}}^T\mathbf{Y}_c$.

3: $\mathbf{E} = \mathbf{Y}_s^T\mathbf{H}_s\mathbf{A}\left(\mathbf{A}\mathbf{H}_s^T\mathbf{H}_s\mathbf{A}^T\right)^{-1}$

4: $\mathbf{F}^T = \mathbf{EA}$

Output: The hyperspectral image $\mathbf{F}$

It is easy to see that the computational complexity of Algorithm 12, is dominated by the computation of the inverse matrices in steps 2 and 3. However, the size of these matrices depends

on the size of the low-dimensional subspace which is considered small. Therefore, the inversion of these matrices has computational complexity $\mathscr{O}(P^2 \log P)$.

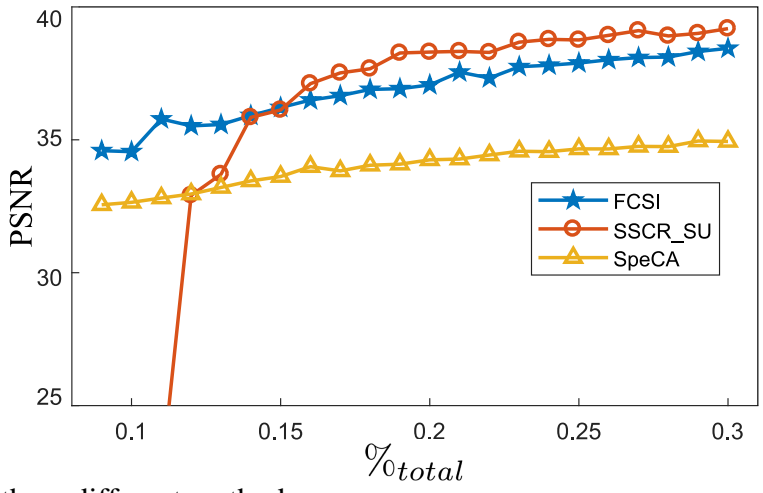
7.1.2.4. Estimation of the size of the low-dimensional subspace (P). The computation complexity of Algorithm 12 depends on the dimension of $\mathscr{S}$ i.e. P. For this reason, it is necessary to estimate the value of P in order to reduce the reconstruction time. Since $K_c, K_s \geq P$ and the sensing matrices $\mathbf{H}_s \in \{0,1\}^{K_s \times MN}, \mathbf{H}_c \in \{0,1\}^{K_c \times L}$ are independently drawn at random from a Bernoulli (1/2) distribution, the matrices $\mathbf{Y}_c$ and $\mathbf{Y}_s$ still lie in a low dimensional subspace, i.e., the ranks of the sensing matrices are precisely P. Considering the noiseless case, let $\lambda_1(\mathbf{Y}_c) \geq \lambda_2(\mathbf{Y}_c) \geq \cdots \geq \lambda_{K_c}(\mathbf{Y}_c)$ be the eigenvalues of $\mathbf{Y}_c$ sorted in descending order. It is easy to see that the P eigenvectors corresponding to the largest eigenvalues, span the range of $\mathbf{Y}_c$, and $\lambda_{P+1}(\mathbf{Y}_c) \geq \cdots \geq \lambda_{K_c}(\mathbf{Y}_c)$ are approximately zero. Thus, the dimensionality of $\mathscr{S}$ is given by

$$\tilde{P} = \arg\max_j \frac{\lambda_j(\mathbf{Y}_c) - \lambda_{j+1}(\mathbf{Y}_c)}{\lambda_{j+1}(\mathbf{Y}_c)}, \tag{157}$$

which finds the maximum gap between two adjacent eigenvalues using the relative error. Notice that this procedure can be carried out with the matrix $\mathbf{Y}_s$ and the same results will hold. This is because, as for the matrix $\mathbf{Y}_c$, $\lambda_{P+1}(\mathbf{Y}_s) \geq \cdots \geq \lambda_{K_c}(\mathbf{Y}_s)$ are approximately zero since $\mathbf{Y}_s$ lies in a subspace of rank P.

7.1.2.5. Simulations and Results. In this experiment, the 145×145 pixel Indian Pines image is used as the test image for this experiment. A total of 20 water absorption and noisy bands were removed from the original 220 bands, leaving 200 spectral bands for the experiment

Figure 42. Quality of the reconstruction measured in PSNR vs total compression ratio



Note: We evaluate three different methods

Zhang et al. (2016). The estimation of P for all algorithms was fixed as $\tilde{P} = 14$, which is the result of applying Hysime Bioucas-Dias and Nascimento (2008) directly on $\mathbf{F}$. The noise considered in the experiment is given by the noise of the image and the one obtained by imposing a low range constraint. In Martín and Bioucas-Dias (2016) the the spectral compressive acquisition (SpeCA) shows that it is enough to choose $n_v = MN$, $m_b = 1$ and the remaining compression is used for m_a. Similarly, the spatial–spectral compressed reconstruction based on spectral unmixing (SSCR_SU) Wang et al. (2018c) shows good behavior when choosing $N_{spa} = 0{,}04MN$ and the rest of the total compression is used for N_{spe}. For this reason, in these simulations, the configuration of the total compression rate was established following their respective suggestions.

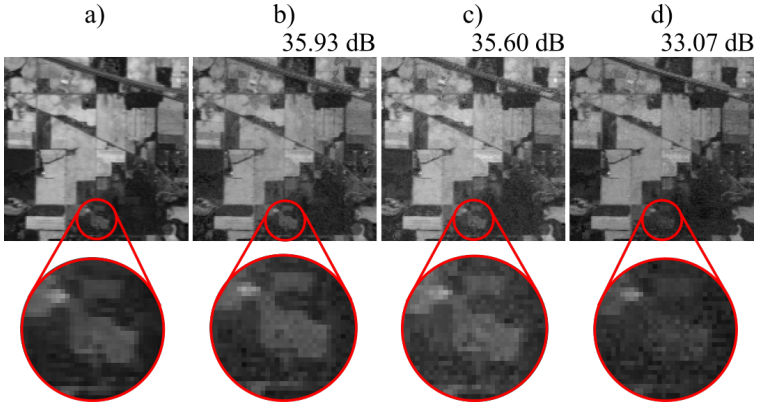
Numerical results for the quality of the reconstruction are shown in Figure 42 where the total compression rate varies from 0.09 to 0.3 in steps of 0.01. Notice that the proposed method performs better when the compression ratio is low. However, when the number of measurements

increases, the proposed method is similar to those obtained with SSCR_SU algorithm. It is worth noting that due to the structure of Φ_{spa}, which randomly select spectral pixels extracting from the

Table 6. Reconstruction time for different data sets and compression ratio.

Reconstruction Time [s]			
Indian Pines			
$\%_{total}$	FCSI	SSCR_SU	SpeCA
.10	**0.04**	2.04	47.83
0.15	**0.05**	2.06	48.10
0.20	**0.05**	1.98	48.98
0.25	**0.07**	0.86	51.56
0.30	**0.15**	0.28	84.46
Pavia University			
P	FCSI	SSCR_SU	SpeCA
5	**2.65**	13.59	58.58
10	**2.38**	14.12	155.71
15	**2.47**	13.43	386.44
20	**2.51**	14.39	717.73
30	**2.60**	15.28	997.36

Figure 43. Indian Pines in band 112

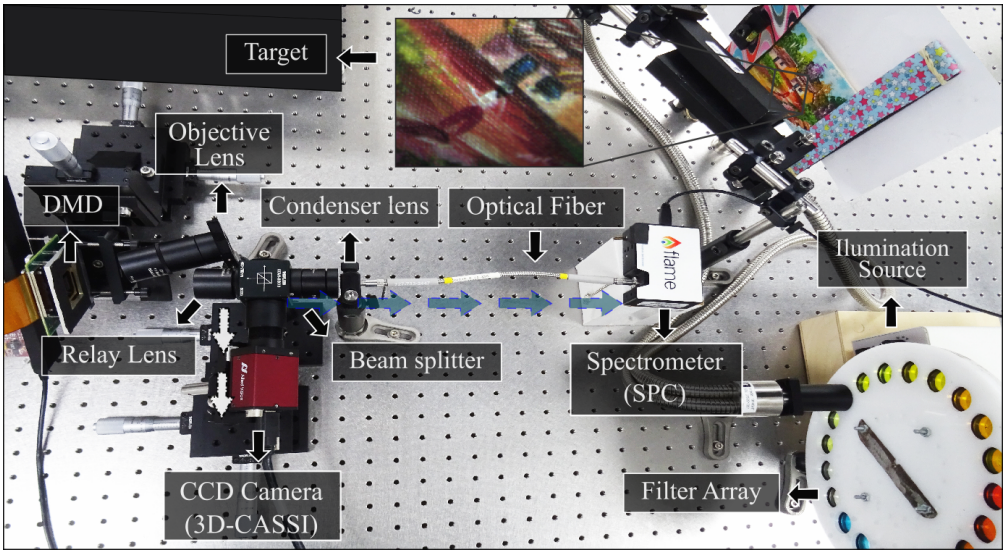


Note: a) original band, Reconstructed band from 15 % of measurements with b) FCSI c) SSCR_SU and d) SpeCA.

full image, SSCR_SU requires at least one pure signature per feature to be sensed. Additionally, the time needed for each of the reconstruction methods is summarized in Table 9. It can be seen that the proposed method is the fastest in comparison with the other methods. Specifically, the proposed method is up to 5.6 times faster than SSCR_SU and 563 times faster than SpeCA. To visualize the reconstructions, Figure 43 shows the Indian Pines dataset in band 112 and the same reconstructed band using 15 % of measurements for all methods. Note that the proposed method in the zoomed version is much cleaner than its counterparts.

7.1.2.6. Validation in a Real Testbed Implementation. The testbed shown in Figure 44 was used to implement the SPC and emulate the 3D-CASSI scheme to verify the proposed al-

Figure 44. Test-bed implementation of the fusion of SPC and 3D-CASSI scheme.

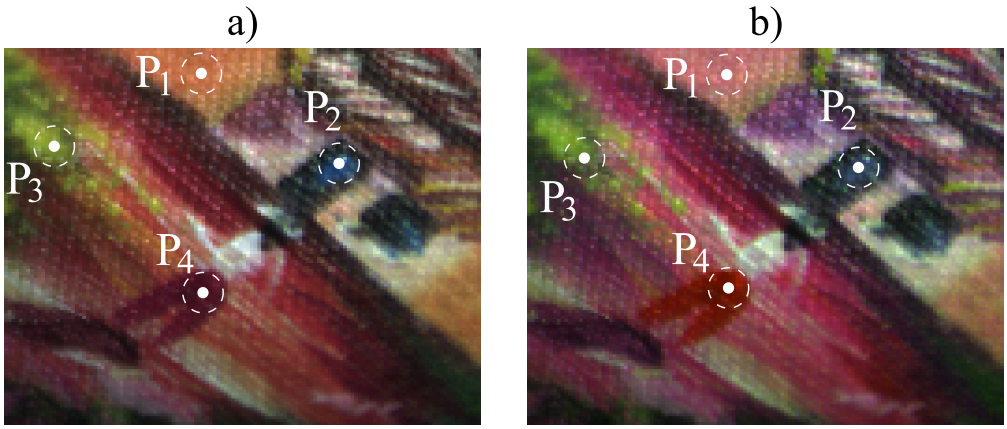


Note: In the SPC arm, the target is illuminated by a lamp, where the objective lens focuses the scene in the DMD, then, the encoded light is condensed in a single point where is sensed by the spectrometer. In the 3D-CASSI scheme, the filter array is used to emulate the spectral bands and similar to the SPC arm the target is encoded and then it is sensed with a CCD camera instead of spectrometer.

gorithm in the laboratory. This prototype consists of two main systems. The SPC arm is composed of a 100*mn* objective lens; a digital micro-mirror device (DMD) used as a spatial light modulator; a 100*mm* relay lens; an F220SMA-A as a condenser lens which is connected to an Ocean Optics Flame S-VIS-NIR-ES spectrometer through an optical fiber. The target in this arm is illuminated by a lamp for the visible spectrum. The 3D-CASSI arm is composed of a set 18 of filters, each with a bandwidth of 13*nm* approximately, between 458*mn* and 638*mn*; the same objective lens relay and a CCD camera are used. A beam splitter is used to ensure that both optical systems observe the same target. To obtain the compressive measurements in both architectures, the 3D-CASSI scheme was first captured. For this, each coding pattern filter was emulated using the filter array and the DMD as explained in detail in Rueda et al. (2015a). To guarantee the reorder process presented in (145), the number of shots and emulated coding pattern filters must be the same. Specifically, for this experiment, a total of 8 shots were acquired, and each shot guarantees that a different coding pattern filter is emulated in each spatial position. So, that, at the end of the shots, each filter has encoded all the pixels and, therefore, the reordering process can be carried out. On the other hand, white illumination was employed to obtain the SPC measurements, and the DMD was randomly configured at each shot to block 50 % of the light.

In order to validate the testbed, the target was sensed with $M = N = 128$, and 72 spectral bands of the scene are reconstructed. It should be clarified that the bands in the spectrometer match the spectral range of the filter array. In these cases, the compression ratio used was established for both architectures as $\%_{3D} = 0{,}11$ and $\%_S = 0{,}08$ for the 3D-CASSI scheme and SPC, respectively. Figure 45 shows the RGB version of the target and the false-color obtained with the reconstructed

Figure 45. a) RGB image b) False color obtained from the reconstructed spectral image.

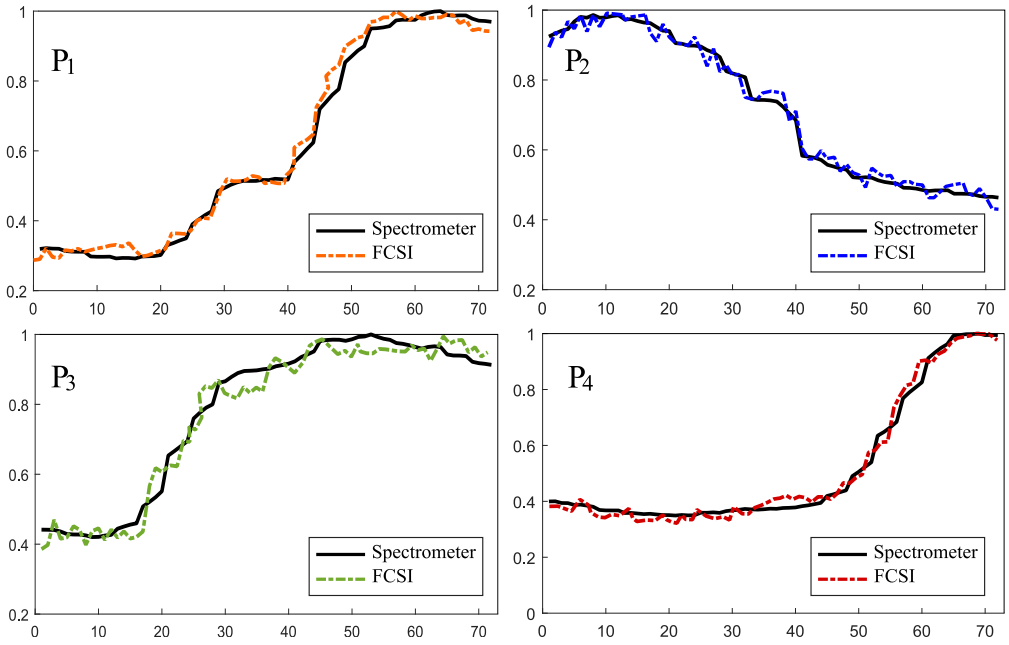


scene with $P = 6$. It can be seen that the spatial information of the scene is preserved with the proposed method. The spectral reconstruction accuracy of the proposed method was validated by choosing four points in the scene and then measuring them with the spectrometer. Figure 46 shows the spectral points obtained with the spectrometer and the reconstructed method. Notice that the proposed method provides noisier signatures; however, it follows the reference spectra from the spectrometer. This result verifies the applicability of the proposed approach to reconstruct images from the fusion of these two types of measurements.

7.2. Coded Aperture Design for High Level Task

The idea behind these proposed methods is to employ high-level tasks directly from the compressive domain bypassing the signal recovery stage. This thesis focuses on subspace clustering applications and a deep learning approach, which can be generalized to high-level tasks like classification and semantic segmentation. Furthermore, this thesis shows that the quality of the high-level task directly depends on the CA, which can be designed to improve the performance of the task.

Figure 46. Spectral Signatures of P_1, P_2, P_3 and P_4 of the target with the spectrometer and the proposed method.



7.2.1. *Coded Aperture Design for Compressive Spectral Subspace Clustering Hinojosa et al. (2018).* This technique is based on the 3D-CASSI scheme present in Section 7.1.2.2. To remember, the sensing process of this architecture can be represented as

$$\mathbf{Y} = \Phi\mathbf{F}, \tag{158}$$

where $\mathbf{Y}$ is the compressive measurements, and Φ represent the coded aperture.

7.2.1.1. Coding Pattern Design. Recent works in CSI have focused on properly designing the coding patterns in order to reconstruct the underlying spectral scene better citearguello2014colored,arguello2013rank. These designs use the restricted isometry property (RIP) as the main optimization criterion. On the other hand, because this thesis aims to perform classification

on the compressed measurements, the design of the coding patterns must preserve the similarity among the spectral signatures. In order to design the coding patterns matrix ■, the following three design criteria are considered.

7.2.1.2. Sensing Scheme. The entries of the matrix Φ are chosen from a Bernoulli distribution $(\Phi)_{s,k} \sim Be(p)$. Therefore, the entries of the s-th coding pattern can be expressed as

$$(\phi^s)_k = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } q, \end{cases} \tag{159}$$

for $k = 0, 1, \cdots, L-1$, where $q = 1-p$. A projection matrix with this structure simply carries out a random sampling on the data vectors, across all the spectral bands, before performing element-wise addition. Considering that surface-emitted spectral signatures are, in general, relatively smooth functions of wavelength Gu et al. (2000), acquiring the information of different sets of adjacent spectral bands at each snapshot will preserve the original signal structure. Therefore, the intuition is to randomly sample neighboring spectral bands instead of randomly sampling all the spectral data vectors, which could add outliers to the measurements.

For each coding pattern ϕ^s, randomly select two cutoff wavelengths $\lambda_{k_1}, \lambda_{k_2} \in \{0, 1, \cdots, L-1\}$ such that $\lambda_{k_1} < \lambda_{k_2}$ and $\lambda_{k_2} - \lambda_{k_1} + 1 = \Delta$, where Δ is defined as the coding pattern bandwidth. Then,

the band-structured random matrix can be expressed as

$$(\phi^s)_k = \begin{cases} 1, & \text{with prob. } \frac{1}{2} \iff \lambda_{k_1} \leq k \leq \lambda_{k_2} \\ 0, & \text{otherwise.} \end{cases} \tag{160}$$

Equation (160) can be alternatively written as

$$(\phi^s)_k = \delta\left(\lfloor \lambda_{k_1}/k \rfloor\right) \delta\left(\lfloor k/\lambda_{k_2} \rfloor\right) \varphi_k, \tag{161}$$

where $\delta(\cdot)$ is the Kronecker delta function and $\varphi \in \mathbb{R}^L$ is a random vector whose entries follow a Bernoulli distribution with probability $\frac{1}{2}$, i.e., $\varphi_k \sim Be(\frac{1}{2})$.

7.2.1.3. Preserving Similarities. The success of subspace clustering on the compressed measurements depends fundamentally on how the coding matrix Φ affects the mutual similarities of the spectral signatures. A usual measure of similarity among two vectors is the cosine of the angle between them. Then, assuming that the vectors have unit length, the similarity between two compressed measurements $\mathbf{y}_j = \Phi\mathbf{f}_j,\ \mathbf{y}_{j'} = \blacksquare\mathbf{f}_{j'}$ is defined as

$$\text{sim}(\mathbf{y}_j, \mathbf{y}_{j'}) = \mathbf{y}_j^T \mathbf{y}_{j'} = \mathbf{f}_j^T \blacksquare^T \blacksquare \mathbf{f}_{j'} \quad j \neq j', \tag{162}$$

where $\mathbf{y}_j \in \mathbb{R}^S$ and $\mathbf{f}_j \in \mathbb{R}^L$ correspond to the j-th column of the matrices $\mathbf{Y}$ and $\mathbf{F}$, respec-

tively. If the columns of ■ are normalized, it is possible to decompose the matrix ■^{T}■ as

$$\blacksquare^T \blacksquare = \mathbf{I} + \Theta, \tag{163}$$

where

$$\Theta_{kk'} = (\Phi_k)^T \Phi_{k'} \quad k \neq k', \tag{164}$$

Φ_k denotes the k-th column of Φ, and $\Theta_{kk} = 0$. Observe that the matrix Θ collects all the entries outside the diagonal of ■^{T}■. Therefore, if $\Theta_{jj'} = 0 \; \forall j, j'$, the matrix ■^{T}■ would be equal to $\mathbf{I}$ and the similarities of the spectral signatures would be exactly preserved in the compressed measurements. However, because the matrix ■ has more columns than rows, all the entries of Θ could be mostly small but not equal zero Kaski (1998). Considering that a linear mapping such as that in (137) can cause significant distortions in the compressed measurements if ■^{T}■ is not approximately $\mathbf{I}$, the proposed coded aperture design should minimize the entries of Θ.

7.2.1.4. Information Acquisition. In order to better discriminate among the classes, new information from the underlying spectral scene should be acquired in each measurement shot. Therefore, the coding patterns should be linearly independent, i.e., the matrix ■ should be full rank. Additionally, the number of measurements acquired for all spectral bands should be approximately the same, i.e., the matrix ▬^{T} should approximate to the identity matrix $\mathbf{I}$. Specifically, this can be attained by decomposing the matrix ▬^{T} as

$$\blacksquare^T = \mathbf{I} + \Lambda, \tag{165}$$

where

$$\Lambda_{ss'} = \phi^s (\phi^{s'})^T \quad s \neq s', \tag{166}$$

and $\Lambda_{ss} = 0$. Therefore, the minimization of the entries of Λ should be considered in the coded aperture design.

7.2.1.5. Optimization Algorithm for Coding Patterns Design. Taking into account the previous considerations, the proposed coding pattern design can be succinctly expressed as the following optimization problem

$$\begin{aligned} \underset{\{\phi^0,\phi^1,\cdots,\phi^{S-1}\}}{\arg\min} \quad & \|\Theta\|_F^2 + \|\Lambda\|_F^2 \\ \text{subject to} \quad & \Theta = \blacksquare^{\mathbf{T}}\blacksquare - \mathbf{I}, \;\; \Lambda = \blacksquare\blacksquare^{\mathbf{T}} - \mathbf{I}, \\ & \text{Rank}(\blacksquare) = S, \\ & (\phi^s)_k = \delta\left(\lfloor \lambda_{k_1}/k \rfloor\right) \delta\left(\lfloor k/\lambda_{k_2} \rfloor\right) \varphi_k, \end{aligned} \tag{167}$$

for $s = 0, \cdots, S-1$ and $k = 0, \cdots, L$. This optimization problem can be efficiently solved with the procedure summarized in Algorithm 13. Specifically, lines 2 to 4 generate the first filter ϕ^0 , which has a band structure with a predefined bandwidth Δ. Then, lines 6-9 are intended to minimize the number of times a spectral band is sensed. Specifically, the algorithm counts how many spectral bands have been sensed in a certain bandwidth, and then the banded section with less information is chosen (expressed in step 9), complying with the criteria of subsection 7.2.1.4.

Finally, this thesis chooses the position in which the inner products are minimized. This is attained by minimizing the elements outside the diagonal, i.e., by minimizing the sum of the values in the neighborhood (step 12) expressed in steps from 15 to 18. As observed in Fig. 47, a random design of Φ entries may lead to oversampling a subset of spectral bands (green line) while leaving some spectral bands unsampled (red line).

Algorithm 13 Optimal Coding Patterns Design

Input: Number of bands L, number of shots S, bandwidth $\Delta > 0$.

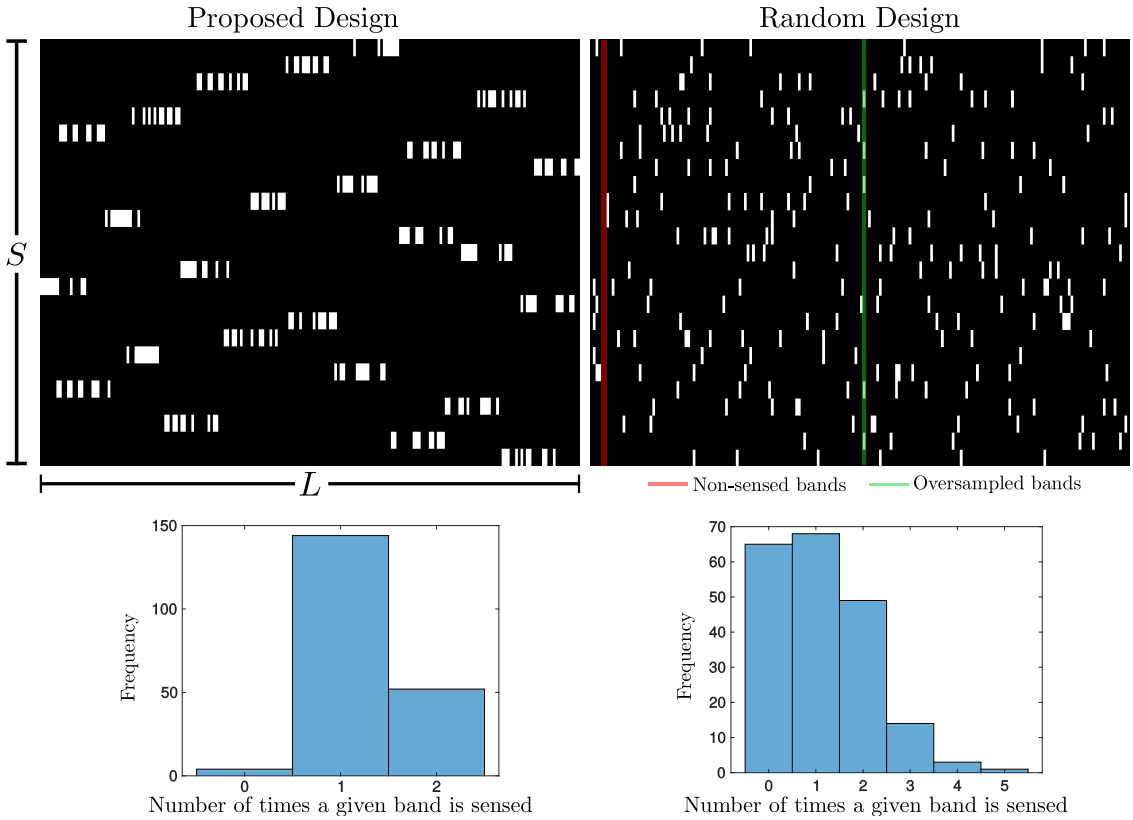
Initialization :

1: $\blacksquare \leftarrow \mathbf{0}_{S,L}$
2: Randomly select $\lambda_{k_1}, \lambda_{k_2}$ such that $\lambda_{k_2} > \lambda_{k_1}$ with $\lambda_{k_2} - \lambda_{k_1} + 1 = \Delta$
3: Select $\varphi \in \mathbb{R}^L$such that $\varphi_k \sim Be(\frac{1}{2})$
4: $\left(\phi^0\right)_k \leftarrow \delta(\lfloor \lambda_{k1}/k \rfloor)\delta(\lfloor k/\lambda_{k2} \rfloor)\varphi_k$
5: **for** $s \leftarrow 1$ to $S-1$ **do**
6: **for** $i \leftarrow 0$ to $(L-\Delta)$ **do**
7: $\mathbf{u}_i \leftarrow \sum_{s'=0}^{s} \sum_{k=i}^{i+\Delta-1} (\phi^{s'})_k$
8: **end for**
9: $\hat{i} \leftarrow \arg\min_i \mathbf{u}_i$
10: $\tilde{\ell} = 0$
11: **for** $i \leftarrow \hat{i}$ to $(\hat{i}+\Delta-1)$ **do**
12: $\mathbf{b}_{\tilde{\ell}} \leftarrow \sum_{s'=0}^{s} \sum_{k=(i-1)}^{i} (\phi^{s'})_k$
13: $\tilde{\ell} = \tilde{\ell} + 1$
14: **end for**
15: **for** $\tilde{i} \leftarrow 0$ to $\lfloor \frac{1}{2}\Delta \rfloor$ **do**
16: $\hat{\ell} \leftarrow \arg\min_{\tilde{\ell}} \mathbf{b}_{\tilde{\ell}}$
17: $(\phi^s)_{\hat{\ell}+\tilde{i}-1} \leftarrow 1$
18: $\mathbf{b}_{\hat{\ell}} \leftarrow \infty$
19: **end for**
20: **end for**

Output: $\blacksquare$

7.2.1.6. Theoretical Results. In the previous section, the optimization algorithm for coding pattern design, proposed in (167), seeks at improving the Φ matrix orthogonality by mini-

Figure 47. Examples of coding patterns generated by the proposed (left) and random (right) design, respectively.



mizing $\|\Theta\|_F^2 + \|\Lambda\|_F^2$. Then it expected that, with high probability, the ℓ_2 norm of $\mathbf{f}_j$ vectors will not change significantly when compressed by the matrix Φ, i.e., $\|\Phi\mathbf{f}_j\|_2^2 \approx \|\mathbf{f}_j\|_2^2$. In this section, using concentration of measure Ledoux (2005), this thesis analyzes the structure of Φ and provide a theoretical bound for such probability. In order to establish a concentration of measure, first observe that the matrix Φ can be decomposed as the product of two matrices as $\Phi = \hat{\Phi}\mathbf{J}$, where $\hat{\Phi} \in \mathbb{R}^{S \times S\Delta}$ is a block diagonal matrix with random vectors φ^s, following a Bernoulli distribution $Be(\frac{1}{2})$, on its diagonals. Fig. 48 depicts the structure for $\hat{\Phi}$ and, $\mathbf{J} \in \mathbb{R}^{S\Delta \times L}$, respectively. The matrix $\mathbf{J}$ can be viewed as a band-subset random selection matrix which constructs vectors partitioned in S blocks

of length Δ, containing the information of Δ random-selected neighboring spectral bands of a particular pixel $\mathbf{f}_j$. Specifically, the block signal is denoted as $\bar{\mathbf{f}}_j \mathbf{J}\mathbf{f}_j = \left[(\bar{\mathbf{f}}_j^0)^T, (\bar{\mathbf{f}}_j^1)^T, \cdots, (\bar{\mathbf{f}}_j^{S-1})^T\right] \in \mathbb{R}^{S\Delta}$, with energy distribution across blocks $\nu \left[\|\bar{\mathbf{f}}^0\|_2^2, \|\bar{\mathbf{f}}^1\|_2^2, \cdots, \|\bar{\mathbf{f}}^{S-1}\|_2^2\right]^T \in \mathbb{R}^S$. Using this notation, the concentration of matrix Φ is presented in the following theorem.

Theorem 7. Assume that $\bar{\mathbf{f}}_j \in \mathbb{R}^{S\times\Delta}$ is a block-partitioned signal with S blocks of size Δ. Let $\hat{\Phi} \in \mathbb{R}^{S\times S\Delta}$ be a block-diagonal random matrix, where each block on its diagonal corresponds to a random vector φ^s drawn independently, whose entries follows a Bernoulli distribution with probability $\frac{1}{2}$. Then, for any $\varepsilon \in (0,1)$

$$\begin{aligned} P(|\|\hat{\Phi}\bar{\mathbf{f}}_j\|_2^2 - \|\bar{\mathbf{f}}_j\|_2^2| > \varepsilon\|\bar{\mathbf{f}}_j\|_2^2) \\ \leq 2\exp\left\{ -C_1 \min\left(\frac{C_2^2\varepsilon^2\|\nu\|_1^2}{\|\nu\|_2^2}, \right.\right. \\ \left.\left. \frac{C_2\varepsilon\|\nu\|_1}{\|\nu\|_\infty}\right)\right\}\left(1-2^{-S}\right), \end{aligned} \tag{168}$$

where C_1 and C_2 are absolute constants.

Demostración. The proof can be found in the published paper Hinojosa et al. (2018). □

The underlying assumption for the success of the SSC algorithm is that the optimization program (described in the next section, see (170)) recovers a sparse subspace representation of each data point, i.e., a representation whose nonzero elements correspond to the subspace of the given point. In Elhamifar and R.Vidal (2013), the authors provide recovery conditions under which, for data points that lie in a union of linear subspaces, the sparse optimization program in (170) recovers

subspace-sparse representation of data points. Particularly, denote $\mathbf{F}_d$ as the matrix containing all data points $\mathbf{f}_j$ from the subspace $\mathscr{S}_d$ with dimension Q_d and, similarly, denote $\mathbf{F}_{-d}$ as the matrix containing data points in all subspaces except $\mathscr{S}_d$. Further, Let $\mathbb{W}_d$ be the set of all full-rank submatrices $\tilde{\mathbf{F}}_d \in \mathbb{R}^{L\times D_d}$ of $\mathbf{F}_d$. From (Elhamifar and R.Vidal, 2013, Theorem 3), if the condition

$$\max_{\tilde{\mathbf{F}}_d\in\mathbb{W}_d} \tilde{\sigma}_{Q_d}(\tilde{\mathbf{F}}_d) > \sqrt{Q_d}\|\mathbf{F}_{-d}\|_{1,2} \max_{d\neq d'} \cos(\theta_{d,d'}) \tag{169}$$

holds, then for every $\mathbf{f}_j$ in the subspace $\mathscr{S}_q$, the ℓ_1-minimization in (170) recovers a subspace-sparse solution [2].

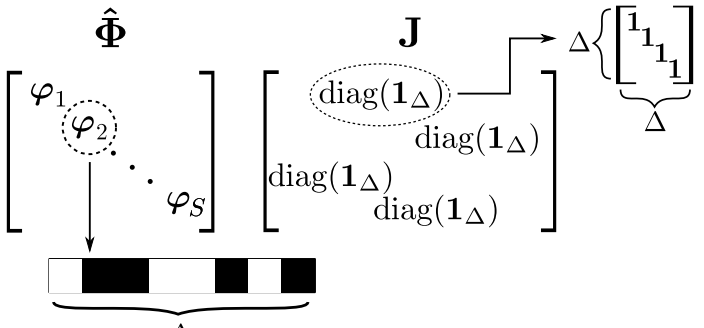
In (169), $\theta_{j,j'}$ is the first principal angle between $\mathscr{S}_d$ and $\mathscr{S}'_d$ and $\tilde{\sigma}_{Q_d}(\tilde{\mathbf{F}}_d) = 1/\|(\tilde{\mathbf{F}}_d^T\tilde{\mathbf{F}}_d)^{-1}\tilde{\mathbf{F}}_d^T\|_{2,2}$ denotes the Q_d-th largest singular value of $\tilde{\mathbf{F}}_d$. Because the matrix Φ preserves the ℓ_2 norm of $\mathbf{f}_j$ with high probability given by (168), then $\tilde{\sigma}_{Q_d}(\tilde{\mathbf{F}}_d)$, which is induced by the ℓ_2 norm, will also be preserved in the compressed domain, i.e., $\tilde{\sigma}_{Q_d}(\tilde{\mathbf{F}}_d) \approx \tilde{\sigma}_{Q_d}(\Phi\tilde{\mathbf{F}}_d)$. In addition, note that because one of the Φ design criteria is to preserve the similarity (a.k.a, cosine of the angle between two vectors), it is expected that $\theta_{j,j'}$ is also preserved. Therefore, it is possible to infer that if the condition in (169) holds for the spectral pixels $\mathbf{F}$, it will also hold for the compressed pixels $\mathbf{Y}$ with high probability.

7.2.1.7. Compressed Sparse Subspace Clustering with Spatial Regularizer . Assuming that compressed pixels of the same land-cover class lie in one independent subspace, subs-

---

[2] The induced norm $\|\mathbf{F}_{-d}\|_{1,2}$, in (169), denotes the maximum ℓ_2-norm of the columns of $\mathbf{F}_{-d}$.

Figure 48. Block diagonal structure of the matrix $\hat{\Phi}$ and the structure of the **J** *matrix.*



Note: In the figure, $\mathrm{diag}(\mathbf{1}_\Delta)$ is a diagonal $\Delta \times \Delta$ matrix whose diagonal values are one.

pace clustering methods can be used in order to separate them into the same group or cluster. In particular, SSC builds the similarity matrix, which describes the data points membership, by finding a sparse representation for each compressed pixel whose nonzero elements ideally correspond to points from the same subspace. Given the designed matrix Φ and the compressive measurements $\mathbf{Y} = \Phi\mathbf{F}$, the SSC sparse representation model is formulated as the following optimization problem:

$$\begin{aligned} \min_{\mathbf{Z},\mathbf{R}} \quad & \|\mathbf{Z}\|_1 + \frac{\lambda}{2}\|\mathbf{R}\|_F^2 \\ \text{s.t.} \quad & \mathbf{Y} = \mathbf{YZ} + \mathbf{R},\ \mathrm{diag}(\mathbf{Z}) = 0,\ \mathbf{Z}^T\mathbf{1} = \mathbf{1}, \end{aligned} \tag{170}$$

where $\mathbf{1}$ is a one-valued vector, $\mathbf{Z} \in \mathbb{R}^{MN \times MN}$ refers to the representation coefficient matrix and the ℓ_1-norm regularization in this formulation suggests that a sparse representation of a data point finds points from the same subspace. The matrix $\mathbf{R}$ stands for the representation error, and the regularization parameter λ for the sparsity trade-off. The constraint $\mathrm{diag}(\mathbf{Z}) = 0$ is used to eliminate the trivial solution of writing a point as an affine combination of itself and the constraint $\mathbf{Z}^T\mathbf{1} = \mathbf{1}$ ensures that it is a case of affine subspaces Elhamifar and Vidal (2009); Elhamifar and

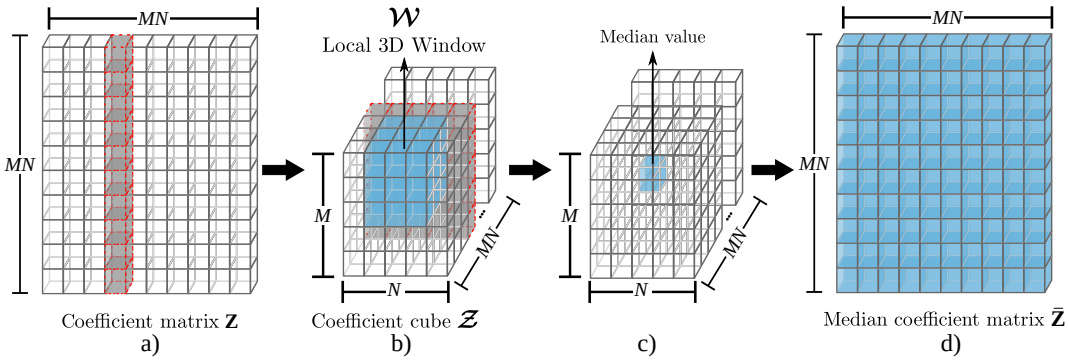
R.Vidal (2013).

Taking into account that neighboring pixels in a spectral image usually consist of similar materials, a smoothing filter can be applied to the sparse coefficient matrix, in order to reduce the representation error, being able to extract more information from the data Zhang et al. (2016). Specifically, the smoothing filter will reduce the noise trying to assign the same representation value to neighboring pixels. In this thesis, such spatial information is effectively incorporated into the similarity matrix by first rearranging the 2-D sparse coefficient matrix $\mathbf{Z} \in \mathbb{R}^{MN \times MN}$ into a 3D cube $\mathscr{Z} \in \mathbb{R}^{M \times N \times MN}$, treating each coefficient vector as a "pixel" in the cube. Unlike Zhang et al. (2016) that perform a 2D average filter in each slide of the cube $\mathscr{Z}$ This thesis proposes to perform the smooth filtering using a 3D median filter with a 3D moving window $\mathscr{W} \in \mathbb{R}^{3 \times 3 \times 3}$. Specifically, $\mathscr{W}$ is moved through $\mathscr{Z}$, on each band, pixel by pixel and replacing each value with the median value of neighboring pixels. Finally, the filtered cube $\mathscr{Z}$ is rearranged back to the matrix $\bar{\mathbf{Z}} \in \mathbb{R}^{MN \times MN}$. This new auxiliary variable is used to regularize $\mathbf{Z}$, hence, the problem of finding a sparse representation coefficient matrix exploiting the spatial information of the scene is formulated as the following optimization program

$$\begin{aligned} \min_{\mathbf{Z},\mathbf{R},\bar{\mathbf{Z}}} \quad & \|\mathbf{Z}\|_1 + \frac{\lambda}{2}\|\mathbf{R}\|_F^2 + \frac{\alpha}{2}\|\mathbf{Z} - \bar{\mathbf{Z}}\|_F^2 \\ \text{s.t.} \quad & \mathbf{Y} = \mathbf{Y}\mathbf{Z} + \mathbf{R},\ \text{diag}(\mathbf{Z}) = 0,\ \mathbf{Z}^T\mathbf{1} = \mathbf{1}, \end{aligned} \tag{171}$$

where α is a regularization parameter denoting the weight of the spatial information in the subspace clustering algorithm. In the subsequent sections, this thesis also refers to the optimization

Figure 49. Visual representation of the median filter step.



Note: a) Sparse Coefficient matrix $\mathbf{Z}$, then it is reshaped as in b) and a median filter is applied to obtain the new values c) and finally it is reshaped to its initial size d).

problem in (171) as S-SSC. The minimization in (171) can be efficiently solved by the alternating direction method of multipliers (ADMM), which is described in detail in the Appendix of the published paper Hinojosa et al. (2018). The solution of (171) corresponds to subspace-sparse representation of the data points, which is used by spectral clustering (SC) to infer the clustering of the data. Specifically, the clustering result is obtained by applying SC to the Laplacian matrix induced by the similarity matrix $\mathbf{W} \in \mathbb{R}^{MN \times MN}$ which is defined as $\mathbf{W} = |\mathbf{Z}| + |\mathbf{Z}|^T$ Elhamifar and Vidal (2009); Elhamifar and R.Vidal (2013). The complete CSI subspace clustering algorithm (CSI-SSC) is summarized in Algorithm 14.

Algorithm 14 Compressive Spectral Subspace Clustering

Input: A set of CSI measurements acquired as $\mathbf{Y} = \Phi\mathbf{F}$, where the coding pattern matrix Φ is obtained with Algorithm. 13.

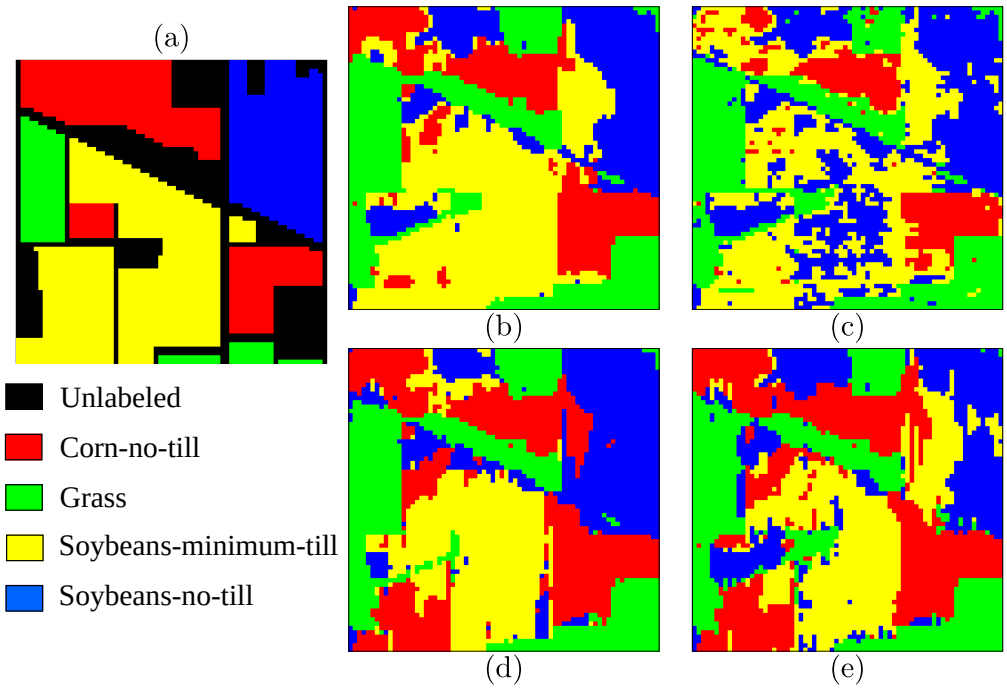
1: Solve the sparse optimization problem in (171) using the ADMM algorithm.
2: Normalize the columns of $\mathbf{Z}$ as $\mathbf{z}_j \leftarrow \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_\infty}$
3: Form a similarity graph representing the data points. Set the weights on the edges between the nodes as $\mathbf{W} = |\mathbf{Z}| + |\mathbf{Z}|^T$.
4: Apply SC Ng et al. (2002) to the similarity graph.

Output: Segmentation of the data: $Y_1, \cdots, Y_\ell$

7.2.1.8. Simulations and Results. In order to validate the clustering performance of the proposed coding pattern design, cluster maps and quantitative results are presented for the two hyperspectral scenes. In all the experiments, the coding patterns were generated for parameters $\Delta = 20$ and $S = 25$. Further, white Gaussian noise with 25 dB of SNR was added to the acquired compressed measurements. In addition, the results obtained with the sparse subspace clustering algorithm (SSC), described in (170), when the complete spectral data cube is used as input (Full-data-SSC), are also shown. Fig. 50 presents the obtained visual clustering results on Indian Pines. The quantitative evaluations corresponding to the accuracy for each class, overall accuracy (OA), average accuracy(AA) and Kappa coefficients are shown in Table 7, where all values are given in percentage. Similarly, Fig. 51 and Table 8 present the visual clustering results and quantitative evaluation on the Pavia University, respectively. In the tables, the optimal value of each row is shown in bold and the second-best result is underlined. From Tables 7 and 8, it can be clearly observed that the proposed clustering approach, using the proposed coding patterns, provides comparable results to applying clustering directly on the full spectral data cube. Furthermore, it is observed from the visual clustering maps that, although the reconstruction is avoided, the results obtained with the proposed coding patterns are very similar to the results obtained with the Full-data. This behavior was expected since the proposed coding patterns approximately preserve the similarities among spectral pixels, as it was theoretically shown in section 7.2.1.6.

7.2.1.9. Clustering Time and Spectral Image Reconstruction. In this section, the effectiveness of applying clustering directly on the compressed domain is evaluated. For this purpo-

Figure 50. Visual clustering results on AVIRIS Indian Pines image.



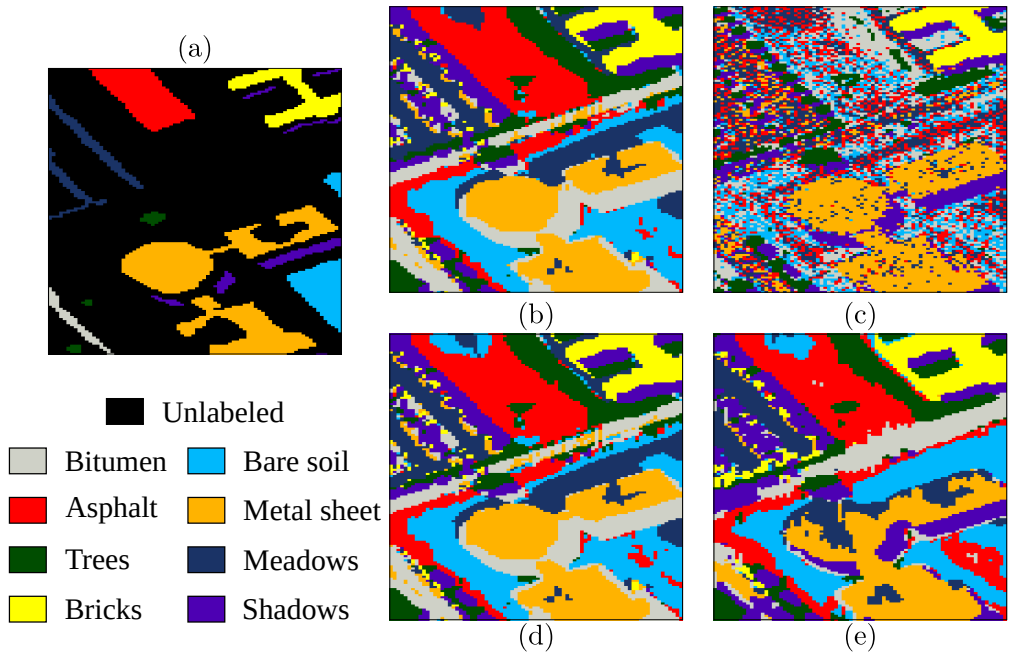
Note: (a) Ground truth. (b) Full-data, (c) Full-data-SSC , (d) Proposed-design and (e) Random-design.

Table 7. Quantitative evaluation of the different clustering results for the AVIRIS Indian Pines Image.

Class	Random-design	Proposed-design	Full-data-SSC	Full-data
Corn-no-till	**73.13**	70.45	48.96	66.77
Grass	95.25	**100**	98.60	**100**
Soybeans-no-till	52.87	**88.80**	70.63	69.54
Soybeans-minimun-till	55.29	60.52	59.23	**80.05**
OA	63.83	73.07	62.62	**76.16**
AA	69.14	**79.94**	69.35	79.09
Kappa	49.26	62.65	47.58	**65.89**

se, a sub-image of ROSIS Pavia university datasets with the size of 64×64 pixels, which contains four land-cover classes: asphalt, meadows, trees, and bricks, were used, see Fig. 52(a). CSI measurements were acquired using the random and proposed coding patterns. Then, 300 iterations of the gradient projection for sparse reconstruction algorithm (GPSR)Figueiredo et al. (2007) were used

Figure 51. Visual clustering results on ROSIS Pavia University image.



Note: (a) Ground truth. (b) Full-data, (c) Full-data-SSC , (d) Proposed-design and (e) Random-design.

Table 8. Quantitative evaluation of the different clustering results with the AVIRIS Pavia University Image.

Class	Random-design	Proposed-design	Full-data-SSC	Full-data
Bitumen	18.60	88.37	0	**90.70**
Asphalt	71.37	67.25	33.84	**80.26**
Trees	90.38	88.46	**100**	90.38
Bricks	**100**	99.68	99.68	99.68
Bare Soil	46.78	61.40	36.26	**66.67**
Metal sheet	82.90	**97.73**	91.00	**97.73**
Meadows	91.16	**100**	55.02	**100**
Shadows	**99.48**	24.35	98.45	24.35
OA	78.72	83.81	71.45	**86.58**
AA	75.09	78.41	64.28	**81.22**
Kappa	72.63	78.89	62.95	**82.50**

to reconstruct the underlying spectral scene. Fig. 52 presents the obtained visual clustering results on the selected sub-image. In Table 9, the time, quality of the reconstruction, and the result of clustering for the types of coding patterns are shown. From this table, it is possible to observe that the

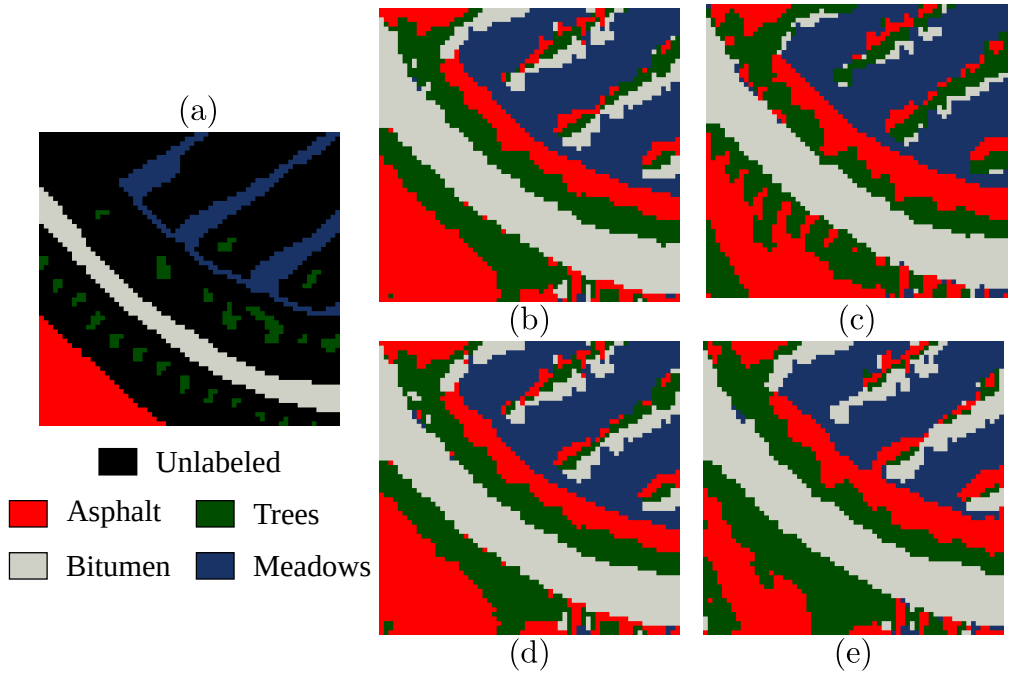
Table 9. Time and classification accuracy when clustering the reconstructed spectral image and the CSI measurements

	Reconstruction		**No Reconstruction**	
	Random Patterns	Proposed Patterns	Random Patterns	Proposed Patterns
		CSI Recovery		
PSNR [dB]	27.92	**34.38**	-	-
Time [s]	**28.56**	26.23	-	-
		Subspace Clustering		
Asphalt	**100**	**100**	**100**	**100**
Meadows	71.65	**99.07**	69.78	97.76
Trees	56.05	63.06	18.85	**65.35**
Bricks	88.04	**99.24**	98.68	98.83
OA	83.43	**94.88**	81.70	94.65
AA	78.94	90.34	71.89	**90.45**
Kappa	77.35	**92.90**	74.74	92.60
Solving (171) [s]	16.62	**15.78**	16.28	10.15
SC Time [s]	118.25	**101.65**	106.11	93.55
Total Time [s]	163.43	143.66	122.39	**103.70**

proposed coding pattern design shows a gain of up to 6 dB in terms of peak signal-to-noise ratio (PSNR) compared to the random patterns. Further, it can be seen that the designed coding pattern improves not only the reconstruction quality but also the clustering result for the two scenarios, i.e., when the subspace clustering is applied after reconstruction and when it is applied directly on the compressed data. Note that the total clustering time of the reconstructions is greater than the time of directly applying clustering on the compressed measurements because it takes into account the reconstruction time. In the simulations, when using the proposed coding patterns, the total clustering time of the reconstructed spectral image was 143.66 [s], while applying clustering directly on the compressed measurements takes only 103.70 [s], obtaining very similar classification results.

7.2.2. ***Deep Coded Aperture Design: An End-to-EndApproach for Computational Imaging Tasks*.** This method intends to cover a range of CSI linear systems. Therefore, in this

Figure 52. Visual clustering results on a 64×64 *region of Pavia University.*



Note: (a) Ground truth. (b), (c) clustering results on reconstructed images using the random and proposed coding patterns, respectively. (d),(e) clustering results by directly using the compressed measurements acquired with the random and proposed coding patterns respectively.

section, some CSI forward measurements are generalized for one snapshot as

$$\mathbf{g} = \mathbf{H}_{\Phi}\mathbf{f} + \eta, \tag{172}$$

where $\mathbf{g} \in \mathbb{R}^m$ denotes the projected encoded measurements, $\mathbf{f} \in \mathbb{R}^n$ denotes the underlying scene, $\mathbf{H}_{\Phi} \in \mathbb{R}^{m \times n}$ models the sensing matrix whose structure is determined by the setup of the optical coding system and the corresponding CA (Φ), and $\eta \in \mathbb{R}^m$ stands for the noise. To highlight, the CA is the main customizable physical element in an optical coding system, so that, it highly determines the performance of the CI task when using projected encoded measurements. Further, some

optical coding systems enable the acquisition of multiple snapshots of the same scene, assuming that the scene remains constant over a time-lapse, by easily varying the used CA. This process is referred to as multishot acquisition Duarte et al. (2008).

Mathematically, for a number of S snapshots, each projected encoded measurement $\{\mathbf{g}^s\}_{s=1}^S$ is obtained with a different sensing matrix $\{\mathbf{H}_{\Phi^s}\}_{s=1}^S$, modeled as in (172). The multishot acquisition process can be compactly expressed as

$$\tilde{\mathbf{g}} = \tilde{\mathbf{H}}_{\tilde{\Phi}}\mathbf{f} + \tilde{\eta}, \tag{173}$$

where $\tilde{\mathbf{g}} = [(\mathbf{g}^1)^T, \cdots, (\mathbf{g}^S)^T]^T$ stacks the projected encoded measurements, $\tilde{\mathbf{H}}_{\tilde{\Phi}} = [(\mathbf{H}_{\Phi^1})^T, \cdots, (\mathbf{H}_{\Phi^S})^T]^T$ stacks the corresponding sensing matrices, $\tilde{\Phi} = \{\Phi^s\}_{s=1}^S$ is a set containing the corresponding CAs for each snapshot, and $\eta = [(\tilde{\eta}^1)^T, \cdots, (\tilde{\eta}^S)^T]^T$ stacks the measurements noise. The ratio between the amount of observed measurements and the size of the underlying scene is known as compression ratio given by $\gamma = Sm/n$.

The proposal aims to couple the design of the sensing matrix $\tilde{\mathbf{H}}_{\tilde{\Phi}}$ together with the achievement of a CI task of interest using an E2E approach. Thus, it jointly optimizes the set of CAs $\tilde{\Phi}$, and the parameters θ, of a chosen DNN $\mathscr{M}_\theta(\cdot)$, by minimizing a cost function composed by the sum of the loss function $\mathscr{L}_{task}(\cdot,\cdot)$, to achieve the task, and a customizable regularization function $R(\tilde{\Phi})$, that promotes particular properties in the set of CAs. Mathematically, given a set of K scenes $\{\mathbf{f}_k\}_{k=1}^K$, and their corresponding task outputs $\{\mathbf{d}_k\}_{k=1}^K$, the proposed coupled optimization

problem is formulated as

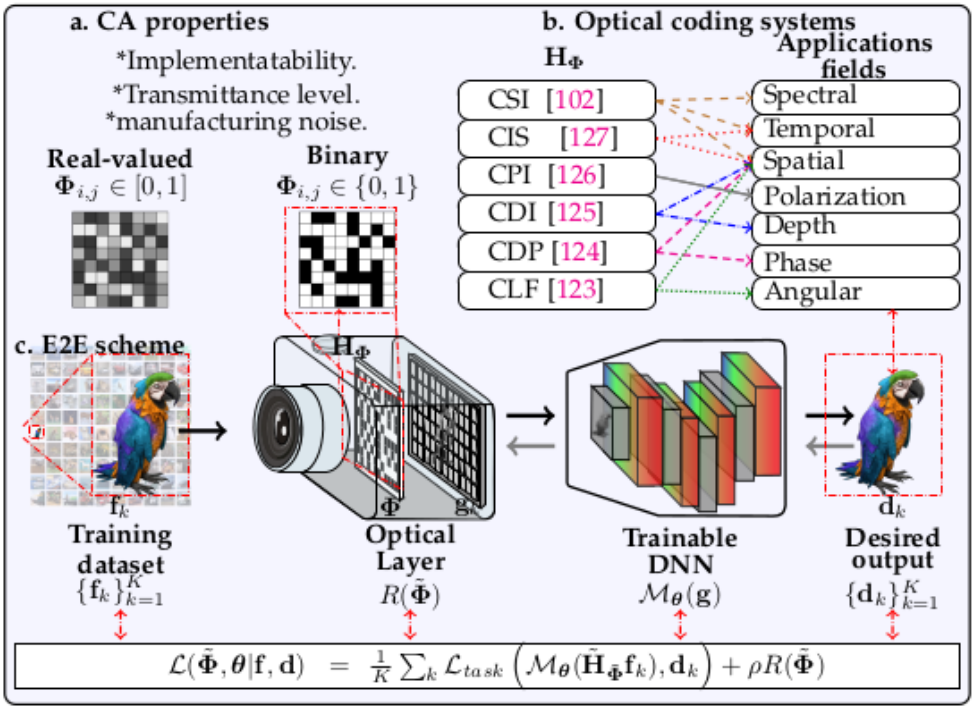
$$\{\tilde{\Phi}^*, \theta^*\} \in \underset{\tilde{\Phi},\theta}{\arg\min}\ \mathscr{L}(\tilde{\Phi}, \theta | \mathbf{f}, \mathbf{d}), \tag{174}$$

$$\mathscr{L}(\tilde{\Phi}, \theta | \mathbf{f}, \mathbf{d}) = \frac{1}{K}\sum_k \mathscr{L}_{task}\left(\mathscr{M}_\theta(\tilde{\mathbf{H}}_{\tilde{\Phi}}\mathbf{f}_k), \mathbf{d}_k\right) + \rho R(\tilde{\Phi}),$$

where, $\rho > 0$ is a regularization parameter.

Then, this thesis follows a two-module procedure that models the encoder and decoder steps

Figure 53. Proposed E2E Approach. (i) The sensing protocol is modeled as a learnable optical layer whose trainable parameter is the CA.



Note: A set of scenes passes through the optical layer to obtain the projected measurements that enter to the hidden convolutional layers up to the loss function to achieve the specific task. The error propagates back up to the CA.

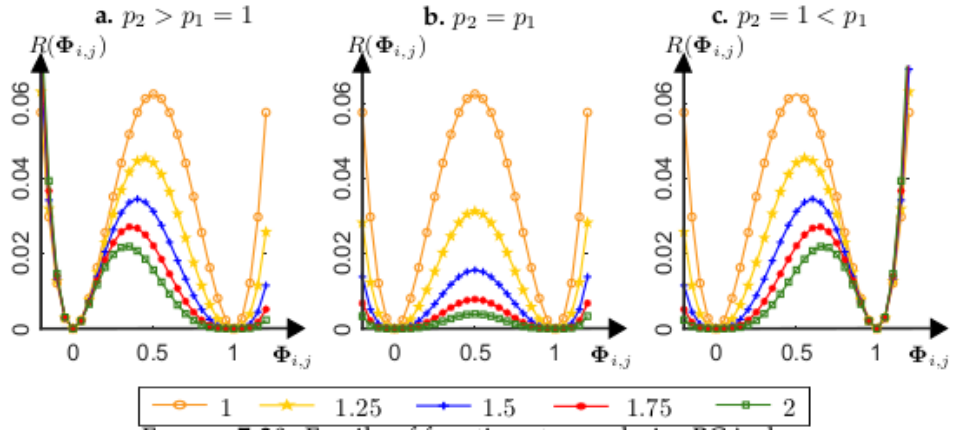
to solve (174). The first, referred to as the sensing protocol module, consists of an optical layer that learns the sensing matrix, $\tilde{\mathbf{H}}_{\tilde{\Phi}}$, with the set of CAs $\tilde{\Phi}$, as the trainable parameters. Notice that $\tilde{\Phi}$ is regularized by the function $R(\tilde{\Phi})$, which aims to guarantee the formation of CAs that satisfy specific implementability, transmittance, number of snapshots, the correlation between snapshots, and conditionality constraints. The optical layer is directly connected to the second module, referred to as the task module, that consists of a chosen trainable DNN with various hidden layers to conclude a task. Figure 53 outlines the coupled E2E proposed approach. It can be seen that the CAs directly affect the task and vice-versa since the backward step to update the values of $\tilde{\Phi}$ takes into account the trade-off between the error given by the loss of the task, and the regularizer function of the CAs.

Remark that once the set of CAs is optimized for the task, the designed sensing matrix $\tilde{\mathbf{H}}^{*}_{\tilde{\Phi}}$ can be used to acquire new projected encoded measurements ($\bar{\mathbf{g}}$), and the pre-trained DNN ($\mathscr{M}_{\theta^*}$) be applied to estimate the task output $\bar{\mathbf{d}}$, as $\bar{\mathbf{d}} = \mathscr{M}_{\theta^*}(\bar{\mathbf{g}})$. Thus, the pre-trained DNN is used as an inference operator.

Next subsections detail the constraints for designing CAs of a coding sensing protocol with the proposed regularizers.

7.2.2.1. Binary Coded Aperture Implementation Constraint. The binarization constraint is addressed by proposing a family of functions whose minima are obtained uniquely when the elements of the CAs are either (0) or (1). Then, it is incorporated in (174) as the regularizer

Figure 54. Family of functions to regularize BCA along different values of the tuple p_1, p_2.



Note: The legends refer to the value of p_2 in **(a.)**, $p_1 = p_2$ in **(b.)**, and p_1 in **(c.)**.

$R(\tilde{\Phi})$ given by

$$R(\tilde{\Phi}) = \frac{1}{S}\sum_{s}\sum_{i,j,\ell}\left((\Phi^{s}_{i,j,\ell})^2\right)^{p_1}\left((\Phi^{s}_{i,j,\ell}-1)^2\right)^{p_2}, \tag{175}$$

where $p_1, p_2 \in \mathbb{R}_{++}$ are two hyper-parameters that provide variability in the function curve as illustrated in Fig. 54, where

the behavior of the family of functions along three different combinations of the tuple (p_1, p_2) is presented. In particular, if $p_1 = p_2$ the graph is symmetric, which turns in equal speed to converge to any of the minima when minimizing the function. Meanwhile, when $p_2 > p_1$ or $p_1 > p_2$, the function presents a bias that leads to a faster direction to converge to one of the minima, 0 or 1. Introducing p_1 and p_2 to control the bias will be useful to determine the transmittance of the designed CAs.

Further, works in Wagadarikar et al. (2008); Candès and Wakin (2008); Higham et al. (2018)

address the binarization constraint through CAs composed by ± 1 values[3] This thesis addresses this constraint by proposing a family of functions whose minimums are found uniquely when the elements of the CAs are either -1 and 1. Hence, the proposed family of functions is incorporated in (174) as the regularizer $R(\tilde{\Phi})$ given by

$$R(\tilde{\Phi}) = \frac{1}{S}\sum_{s}\sum_{i,j,\ell}\left((\Phi^{s}_{i,j,\ell}+1)^2\right)^{p_1}\left((\Phi^{s}_{i,j,\ell}-1)^2\right)^{p_2}, \tag{176}$$

which generalizes the function presented in Bacca et al. (2020); Higham et al. (2018). Particularly, when $p_1 = p_2 = 1$, (176) produces the function used in Higham et al. (2018) for reconstruction and in Bacca et al. (2020) for classification.

7.2.2.2. Real-valued Coded Aperture Implementation Constraint. The range of admitted values in a real-valued CA ($\Phi \in [0,1]$) is wider than in the binary one, which mathematically results in a more relaxed constraint. However, the wider the range of values in the mathematical design, the harder the fabrication and calibration in the physical system Diaz et al. (2019). This difficulty can be alleviated by fixing a set $\{\kappa_d\}_{d=1}^{D} \in \mathbb{R}$ of different quantization levels to appear in the design. This thesis addresses this real-valued constraint by proposing a family of functions whose minima are obtained uniquely when the elements of the CAs are the fixed quantization

---

[3] Codification with negative values can be physically achieved by estimating and subtracting the mean light intensity from each measurement, which can be achieved with a full-one CA Duarte et al. (2008) For more physical details see Bacca et al. (2020).

levels κ_d. Then, this thesis incorporates it in (174) as the regularizer $R(\tilde{\Phi})$ given by

$$R(\tilde{\Phi}) = \frac{1}{S}\sum_{s}\sum_{i,j,\ell}\prod_{d}\left(\left(\Phi^{s}_{i,j,\ell} - \kappa_d\right)^2\right)^{p_d}, \tag{177}$$

where $\{p_d\}_{d=1}^{D} \in \mathbb{R}_{++}$ are the hyper-parameters associated for each quantization level. Notice that (177) is a polynomial of degree $\prod_d 2p_d$ whose range is always positive and whose roots correspond to the targeted quantization levels in the CA design. Thus, (177) is the generalization of the previous family of functions in (175) and (176) which are the particular cases of 2 target values, i.e., $D = 2$ with $\kappa_1 = 0, \kappa_2 = 1$ for (175), and $\kappa_1 = -1, \kappa_2 = 1$, for (176), respectively.

7.2.2.3. Coded Aperture Transmittance Constraint. The transmittance level is a crucial property that affects the calibration of the optical coding system and determines the proper utilization of the light in the acquired measurements. Therefore, adjusting the transmittance level is an important step to accomplish a task. For instance, in spectral imaging, a high-transmittance level is desired to increase the signal-to-noise power ratio Rueda et al. (2015b), unlike in X-ray tomography, a low-transmittance level is desired to minimize radiation to the objects Mojica et al. (2017). Specifically, varying the transmittance level unbalances the distribution of the quantization levels in the CA and produces an ill-conditioned sensing matrix; in consequence, the performance of the system can be affected. The general family of functions can indirectly control the transmittance level by adjusting the hyperparameter κ_d. Nonetheless, to achieve an exact targeted value, this thesis addresses this transmittance level constraint by proposing the following regularization function $R(\tilde{\Phi})$ that adjusts the transmittance level while affecting the less the possible performance

of the system

$$R(\tilde{\Phi}) = \frac{1}{S}\sum_{s}\left(\frac{\sum_{i,j,\ell}\Phi^{s}_{i,j,\ell}}{MNL} - T_r\right)^2, \tag{178}$$

where $T_r \in [0,1]$ is a customizable hyperparameter that denotes the targeted transmittance level, with 0 and 1 indicating to block or unblock all the incoming light.

7.2.2.4. Number of Snapshots Constraint. The aim of acquiring multiple snapshots is to efficiently increase the amount of observed information related to the properties of the scene improving the task performance. However, the number of snapshots implies a tradeoff between the task performance and the needed time for acquiring and processing more snapshots Baraniuk et al. (2017). Therefore, it is essential to determine the least amount of optimal snapshots S that achieve the highest performance. This thesis proposes to address this number of snapshots constraint by using the following regularizer

$$R(\tilde{\Phi}) := \sum_{s'}^{S'}\sqrt{\sum_{i,j,\ell}\left(\Phi^{s'}_{i,j,\ell}\right)^2}, \tag{179}$$

where $S' \geq S$ is an upper bound for the number of user-determined snapshots, this equation can be seen as the ℓ_1-norm through the snapshot dimension of the ℓ_2-norm of the vectorization of each CA. This formulation is based on the traditional $\ell_{2,1}$-norm applied on matrices, which has been demonstrated to encourage all values in a column to be zero Eldar and Bolcskei (2009); Lohit et al. (2019). Thus, (179) promotes all entries in a CA to be equal to zero, i.e., implying the no acquisition of the snapshots.

Notice that expression in (179) is not differentiable when all values of a CA are zero Lohit

et al. (2019). Hence, this thesis employs the sub-gradient below in the backward step

$$\frac{\partial R(\tilde{\Phi})}{\partial \Phi^{s'}_{i,j,\ell}} = \begin{cases} \frac{\Phi^{s'}_{i,j,\ell}}{\varphi(\Phi^{s'}_{i,j,\ell})}, & \varphi(\Phi^{s'}_{i,j,\ell}) \neq 0 \\ 0, & \text{otherwise,} \end{cases} \tag{180}$$

where $\varphi(\Phi^{s'}_{i,j,\ell}) = \sqrt{\sum_{i,j,\ell}(\Phi^{s'}_{i,j,\ell})^2}$.

Notice that regularizers proposed up to this section are directly related to assembling properties of the CA, such as the implementability of the obtained quantization levels, the adjustment of the transmittance, and the selection of the number of snapshots to be acquired. Unlike, the following two regularizers are proposed in order to achieve a better performance in the solution of the CI task.

7.2.2.5. Multishot Coded Aperture Correlation Constraint. The correlation between the CAs $\{\Phi^s\}_{s=1}^{S}$ used in a multishot acquisition scheme is crucial to increase the observed information of the underlying scene. Specifically, it has been demonstrated in Correa et al. (2015) that the less correlated the CAs, the greater the amount of acquired with fewer snapshots. This thesis addresses the correlation constraint using a generalized function that minimizes the correlation between the S designed CAs. Then, this thesis incorporates it in (174) as the regularizer $R(\tilde{\Phi})$ given by

$$R(\tilde{\Phi}) = \frac{\sum_{i,j,\ell}\left(\prod_s \Phi^s_{i,j,\ell}\right)}{MNL}. \tag{181}$$

Observe that for $S = 2$, expression in (181) results in the numerator part of the Pearson correlation for two CAs Benesty et al. (2009).

7.2.2.6. Data Driven Conditionality Constraint. The conditionality constraint accounts for the sensing matrix structure $\tilde{\mathbf{H}}_{\Phi}$, which should be linearly independent along its rows and columns, to ease its inversion while preserving the features after projection Hinojosa et al. (2018). Authors in Hinojosa et al. (2018); Mejia and Arguello (2018); Mojica et al. (2017) address this aim by imposing the regularizer $\|\tilde{\mathbf{H}}_{\Phi}^{T}\tilde{\mathbf{H}}_{\Phi} - c\mathbf{I}\|_F^2$, where $\mathbf{I}$ denotes the identity matrix for a constant $c \in \mathbb{R}_{++}$. This thesis extends this regularizer by taking advantage of the available data such that the conditionality is improved according to the specific dataset of interest instead of for general acquisition. Hence, the proposal introduces the following regularizer in (174) to promote a data-driven improved conditionality in the sensing matrix

$$R(\tilde{\Phi}) = \frac{1}{K}\sum_{k} \left\| \tilde{\mathbf{H}}_{\Phi}^{T}\tilde{\mathbf{H}}_{\Phi}\mathbf{f}_k - \mathbf{f}_k \right\|_2^2. \tag{182}$$

7.2.2.7. Modeling Considerations.

7.2.2.8. Trainable Parameters. The number of trainable parameters is a key aspect in the efficiency and performance of the proposed E2E approach. Specifically, when the optical layer contains a set of S 2D or 3D CAs with $N \times M$ and $N \times M \times L$ elements, respectively, the number of trainable parameters ascends to SMN and $SMNL$, respectively. This limits its use for large scale scenes because of the expensive computational memory requirements and the potential overfitting problems Goodfellow et al. (2016). This thesis proposes to reduce the number of traina-

ble parameters by adding spatial structure to each CA so that a kernel $\mathbf{Q}^s$ of size $\Delta_n \times \Delta_m \ll N \times M$ is periodically repeated as follows

$$\Phi^s = \mathbf{1} \otimes \mathbf{Q}^s, \tag{183}$$

for $s = 1, \ldots, S$, where $\mathbf{1}$ denotes a matrix of size $\frac{N}{\Delta_n} \times \frac{M}{\Delta_m}$ with all elements equal to 1, and $\otimes$ represents the Kronecker product. Including such periodicity reduces the total number of trainable variables to $S\Delta_n\Delta_m$ and $S\Delta_n\Delta_m L$ for the 2D and 3D representation, respectively. Note that this model formulation enables to train some optical coding systems with small portions of the training images, known as patches, by training directly $\mathbf{Q}^s$ Gelvez and Arguello (2020). In some 3D applications, this training parameter can be further reduced. For instance, in the colored CA, each color pixel can be expressed as a linear combination of $V < L$ fixed optical filters $\{\mathbf{w}^v \in [0,1]^L\}_{v=1}^V$, so that, each element of the 3D kernel can be rewritten as

$$\mathbf{Q}_{i,j,\ell} = \sum_n \mathbf{w}_\ell^n \mathbf{A}_{i,j}^n, \tag{184}$$

where $\mathbf{A}$ denotes the trainable weights of the linear combinations, and in consequence, the number of trainable parameters is reduced to $S\Delta_n\Delta_m V$.

7.2.2.9. Manufacturing Noise. The manufacturing noise represents a problem in the design of optical elements since it can lower the obtained benefits with the design when implemented in a real setup Diaz et al. (2019); Sitzmann et al. (2018). To overcome this problem, an exhaustive calibration process over the real setup is carried out to achieve a performance as reliable as the sensing model used in the simulations. On the other hand, this problem can also be

addressed by refining the noise modeling in the design. Therefore, this thesis aims at considering two sources of perturbations at each step of the forward propagation in the E2E optimization, one into the projected measurements as expressed in (173), which commonly comes from the level of illumination He et al. (2015), and the other into the CA, which commonly comes from the manufactured process Sitzmann et al. (2018). This thesis adds the latter manufacturing noise as follows

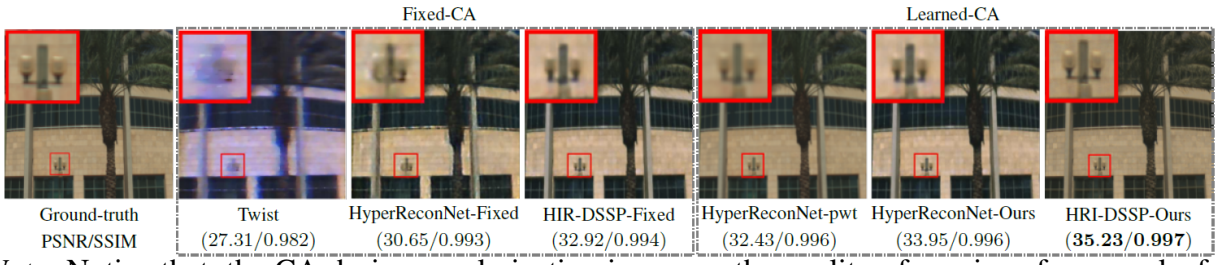
$$\Phi = \Phi + \eta, \tag{185}$$

where $||\eta||_\infty << ||\Phi||_\infty$, and the distribution of the noise η varies according to fabrication processes.

Notice that these regularizes have been designed for specific purposes. However, some of them can be incorporated into the optimization problem to get the desired behavior.

7.2.2.10. Simulation and Results. This experiment aims to show that coupling the design of the sensing protocol and the decoder training increase the quality of CI tasks. For this, this thesis compared the results against the non data-driven recovery method TwIST with TV prior Kittle et al. (2010), and the data-driven recovery networks Hyperspectral Image Reconstruction using a Deep Spatial-Spectral Prior (HIR-DSSP) Wang et al. (2019), and HyperReconNet Wang et al. (2018d). For Twist, this thesis employed a fixed CA. For (HIR-DSSP) and HyperReconNet, different variations are evaluated: first, this thesis evaluated the performance when using a fixed CA generated following a Bernoulli distribution with parameter 0,5, whose results are denoted as HyperReconNet-Fixed, and HIR-DSSP-Fixed. Furthermore, the performance when joi-

Figure 55. RGB mapping comparison of the reviewed data-driven approaches, employing fixed and learned CA into the network.

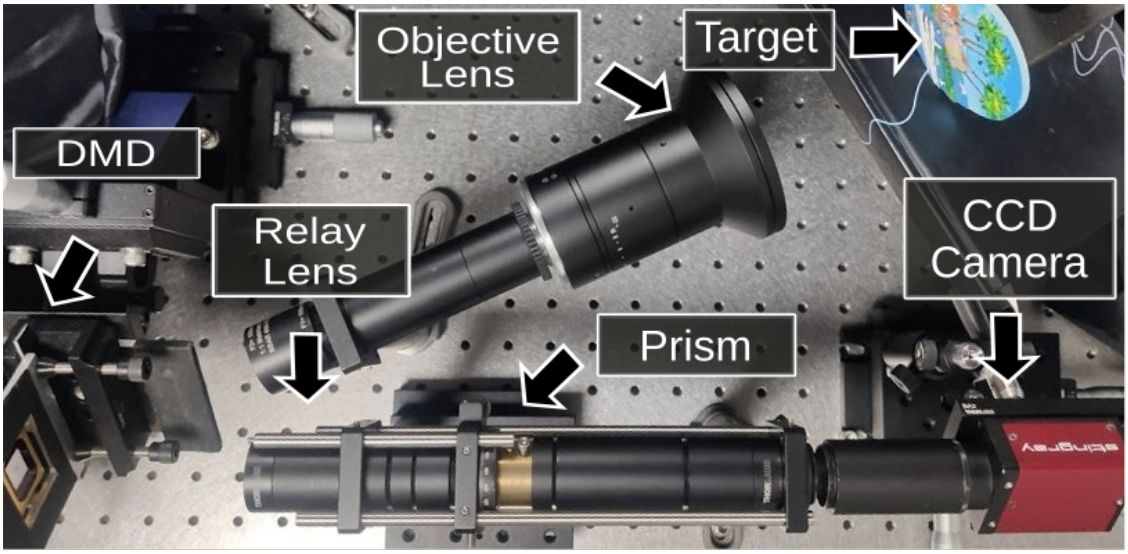


Note: Notice that, the CA design regularization improves the quality of previous frameworks for the recovery task.

ning the design of the CA and the network weights through the proposed binary regularization are evaluated, whose results are denoted as HyperReconNet-Ours, and HIR-DSSP-Ours. Finally, this thesis evaluated the performance when using the strategy proposed by the HyperReconNet Wang et al. (2018d), which learns the CA using a piece-wise threshold (pwt), whose result is denoted as HyperReconNet-pwt. Figure 55 shows an RGB visual representation of the recovered images with the PSNR and SSIM quantitative metrics which show that the proposed method outperforms state-of-the-art methods. Remark the versatility of the proposed regularizers which can be straightforwardly incorporated at any pre-existed net.

7.2.2.11. Validation in a real setup experiment. Section 7.2.2.10 demonstrated that coupling the sensing protocol design and the processing method increases the quality of CI tasks; however, most of the obtained benefits are lowered when applying those methods in real setups Correa et al. (2016b); Bacca et al. (2019). Hence, this experiment validates the E2E approach with a real setup corresponding to one single snapshot of the CASSI testbed laboratory implementation depicted in Fig. 56

Figure 56. Testbed CASSI implementation



Note: the relay lens focuses the encoded light by the DMD into the sensor after dispersed by the prism.

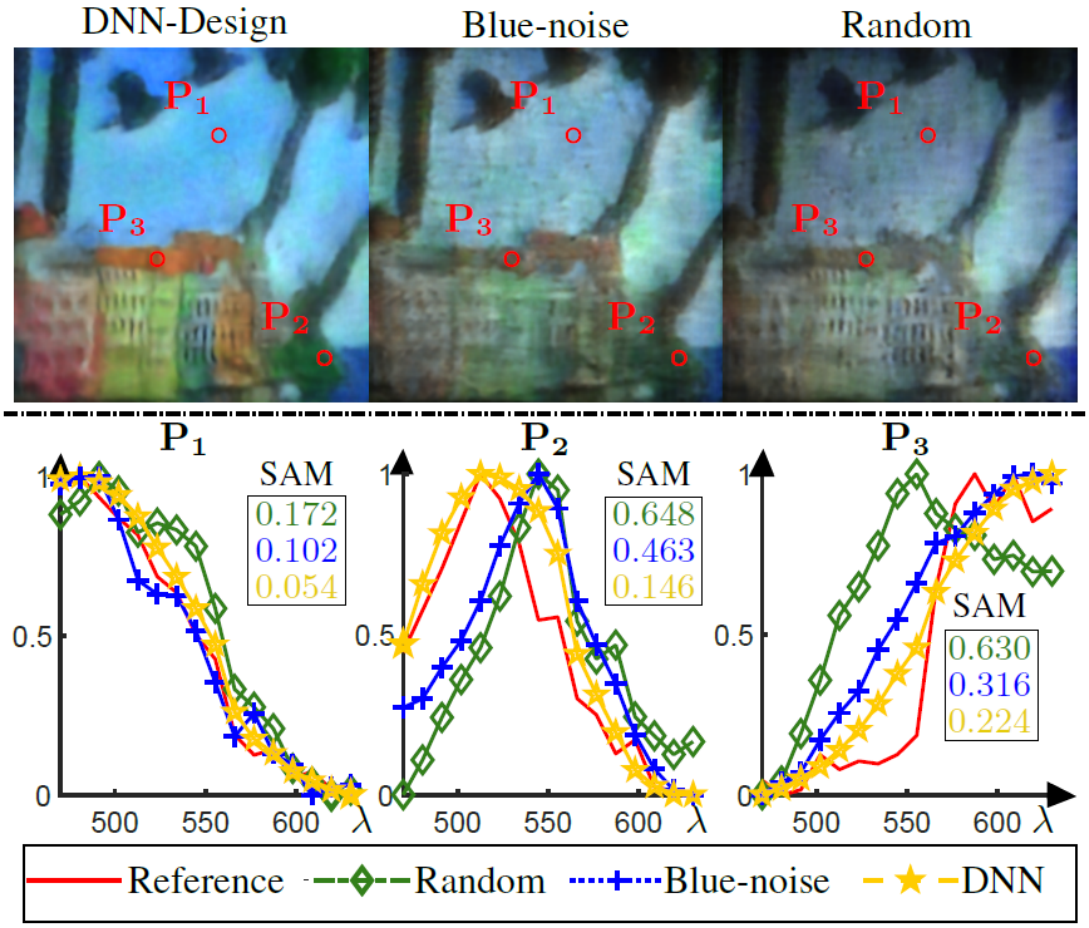
For this, the ARAD dataset was used to train the proposed approach with a compression ratio $\gamma = 0{,}034$. The setup contains a 100-*nm* objective lens, a high-speed digital micro-mirror device (DMD) (Texas Instruments-DLI4130), with a pixel size of $13{,}6\mu m$, an Amici Prism (Shanghai Optics), and a CCD (AVT Stingray F-145*B*) camera with spatial resolution 1388×1038, and pitch size of $6{,}8\mu m$. The 482×512 designed CA is placed at the center of the DMD.

The performance was compared against the results obtained with the same DNN and hyper-parameter tunning process, but with a fixed CA and learning only the network weights, i.e., the main difference is that the optical layer is trainable in the proposed method. For the fixed CA, this thesis used a random CA and a designed blue-noise CA Correa et al. (2016a).

Figure 57 (Top) illustrates an RGB visual representation comparison of the spatial quality between the reconstructions obtained along the three approaches. It can be noticed a significant

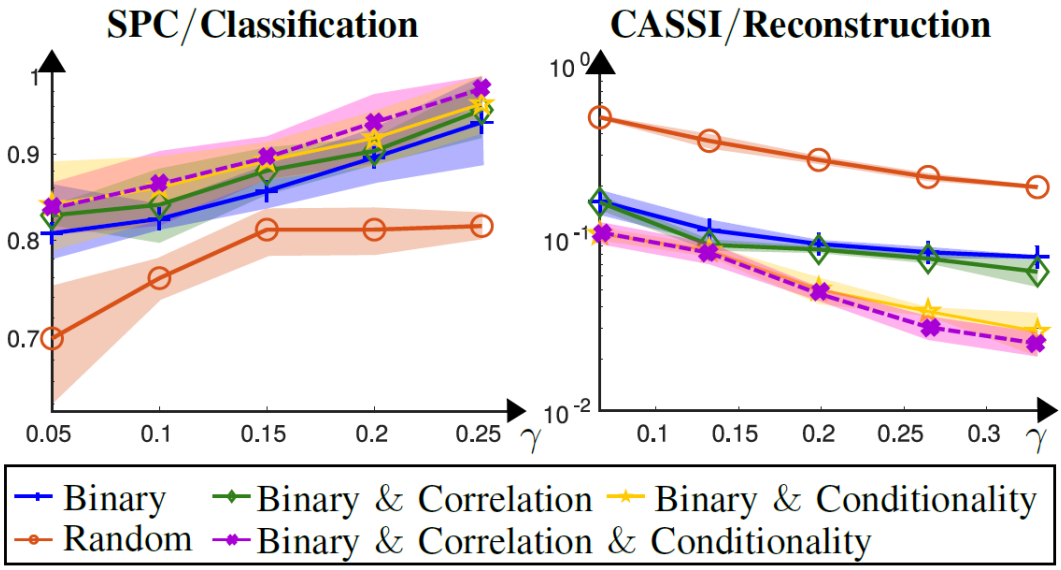
improvement in the visual results for a real setup when coupling the design of the CA in the DNN. Fig. 57 (Bottom) illustrates a comparison of the spectral quality at three random spatial locations whose spectral response was measured in the laboratory with the commercially available spectrometer (Ocean Optics USB2000+). It can be noticed that the coupled approach decreases the spectral angle between the estimated and reference spectral signatures in comparison to the fixed

Figure 57. (Top) RGB visual representation of the three evaluated methods (Net design, Blue-noise and random). (Bottom) Comparison of the normalized spectral signatures at three points in the recovered scenes.



Note: At each point it is shown the quantitative SAM metric to quantify the improvement.

Figure 58. Quality behavior of adding various regularizers.



Note: (Left) Real SPC setup for classification task with the quality measured in terms of accuracy. *(Right)* Real CASSI setup for reconstruction task with quality measured in terms of the SAM metric of the spectral response at three spatial locations.

CAs.

7.2.3. *Quality improvement via regularizers experiment.* This experiment evaluates the effectiveness of binary (175), correlation (181), and conditionality (182) regularizers to guide the solution of the optimization problem to a better quality in the CI task for real setups. The initial value of each regularization was set as $\rho_0 = 1e^{-7}$ except for the binary that was as $\rho_0 = 1e^{-9}$. This thesis used the CASSI and a SPC testbed implementation, where 50 random numbers were printed and used as the target (See supplementary material for further implementation details). Figure 58 compares the task quality across an interval of different compression ratios for five cases: using a random CA; using only binary regularizer; using binary and correlation regularizers; using binary and conditionality regularizers; and using binary, correlation, and conditionality

regularizers. It can be seen that the quality of the task increases as more regularizers are taken into account, such that using the three regularizers together suppressively outperforms the recovery task quality in both setups. This thesis also remarks that the binary regularizer itself provides a significant improvement in comparison to the random scenario.

8. Conclusions and Future works

This dissertation studied the super-resolution phase retrieval problem from coded diffraction patterns. Specifically, the inclusion of a high-resolution coded aperture (CA) allowed the development of a new super-resolution system that was denominated physical super-resolution. The forward models for the different diffraction zones (near, middle, and far), were derived under the assumption that the attainable resolution of the image is determined by the resolution of the CA, assuming that the pixel size of the CA is smaller than the sensor pixel size. Theoretical uniqueness guarantees for all the diffraction zones were provided, establishing that the set of coded apertures has to be designed to increase the probability of a unique solution. Numerical experiments were conducted to evaluate the performance of the super-resolution approach, and it was shown that the reconstruction quality is preserved up to a resolution factor of 4.

From the theoretical results, it was concluded that the CA distribution plays a crucial role in recovering the phase in coded diffraction patterns (CDP). Therefore, this thesis shows two design strategies, one independent of the data, based on a greedy strategy to increase the theoretical recovery probability, and the other based on data, using an end-to-end (E2E) deep learning approach, which is based on modeling a differentiable sensing model of the optical system which is coupled as a layer in a recovery network. Simulations show that both methodologies overcome random codes, but the E2E method shows better results for different coded elements and the three studied diffraction zones.

From the recovery point of view, this thesis proposed a smoothed non-convex least-squares

objective function, which addresses the non smoothness of the traditional phase retrieval formulation. Furthermore, this thesis shows that sparsity prior as total-variation and deep priors included in the proposed smoothing formulation drastically increase the reconstruction quality, resulting in good reconstruction even from a single projection. Besides, the deep unrolling strategy developed in this thesis shows the best performance to recover the phase in a coded diffraction pattern setup, compared with the state-of-the-art methods.

Additionally, mathematical CA design concepts and recovery algorithms were extended to compressive spectral imaging. There, it was shown that the quality significantly increases for different tasks, not only reconstruction but also classification and clustering. Furthermore, these applications can be validated in the optics setups where the performance is maintained.

Future work includes the implementation of the proposed super resolution phase retrieval scheme to validate the obtained results in a real scenario where some physical consideration about the noise and calibration problem needs to be considered. The mathematical concepts developed in this thesis can be extended to other applications that employ CA modulations as compressive seismic data and radar signals.

9. Appendix

Appendix A: Proof of Theorem 3

Since the initialization used in this method, generates a point $\mathbf{x}_0$ near to the real signal $\mathbf{x}$, it suffices to show that the convergence of gradient loops given $\mathbf{x}_0$ lands into the neighborhood of global minimums. To prove Theorem 3, the major step is to prove the following Lemma 7 which characterizes how the error of an estimate decays upon one iteration of Algorithm 2. Once Lemma

7 is established, this thesis takes expectation on both sides of Eq. (76) with respect to i_{t-1}, and apply Lemma 7 one more time to obtain

$$\mathbb{E}_{\{i_{t-1},i_t\}}\left[d_r^2(\mathbf{x}_{t+1},\mathbf{x})\right] \leq (1-\upsilon)^2 d_r^2(\mathbf{x}_{t-1},\mathbf{x}), \tag{75}$$

where $\upsilon \in (0,1)$. This process continues until the initialization point $\mathbf{x}_0$ yields Theorem 3. Then, this thesis focuses on proving Lemma 7 stated bellow.

Lemma 7. Consider the noiseless measurements $q_k = |\langle \mathbf{a}_k, \mathbf{x}\rangle|$ with an arbitrary signal $\mathbf{x} \in \mathbb{C}^n$, and i.i.d $\{\mathbf{a}_k \sim \mathscr{CN}(0,\mathbf{I}_n)\}_{k=1}^m$. If $\alpha \in (0,\alpha_0/n]$ and also $m \geq c_0 n$, then with probability at least $1-2e^{-\varepsilon^2 m/2}$, we have

$$\mathbb{E}_{i_t}\left[d_r^2(\mathbf{x}_{t+1},\mathbf{x})\right] \leq (1-\upsilon) d_r^2(\mathbf{x}_t,\mathbf{x}), \tag{76}$$

for all $\mathbf{x}_t$ satisfying $\frac{d_r(\mathbf{x}_t,\mathbf{x})}{\|\mathbf{x}\|_2} \leq \frac{1}{10}$.

Demostración. Let $\mathbf{h}_t = e^{-j\theta_t}\mathbf{x}_t - \mathbf{x}$ with $\mathbf{x}_t$ and $\theta_t = \arg\min_{\theta\in[0,2\pi)} \|e^{-j\theta}\mathbf{x}_t - \mathbf{x}\|_2$. Then, by definition of $d_r(\cdot,\cdot)$ we have that

$$d_r^2(\mathbf{x}_{t+1},\mathbf{x}) = \min_{\theta\in[0,2\pi)} \|e^{-j\theta}\mathbf{x}_{t+1} - \mathbf{x}\|_2^2 \leq \|e^{-j\theta_t}\mathbf{x}_{t+1} - \mathbf{x}\|_2^2. \tag{77}$$

From Eq. (187) it can be obtained that

$$\mathbb{E}_{k_t}\left[d_r^2(\mathbf{x}_{t+1},\mathbf{x})\right] \leq \mathbb{E}_{k_t}\left[\|e^{-j\theta_t}\mathbf{x}_{t+1} - \mathbf{x}\|_2^2\right]. \tag{78}$$

Notice that

$$\begin{aligned}
&\mathbb{E}_{k_t}\left[\left\|e^{-j\theta_t}\mathbf{x}_{t+1}-\mathbf{x}\right\|_2^2\right] \\
&=\mathbb{E}_{k_t}\left[\left\|e^{-j\theta_t}\left(\mathbf{x}_t-\alpha\left(\mathbf{a}_{k_t}^H\mathbf{x}_t-\vartheta_{k_t}\right)\mathbf{a}_{k_t}\right)-\mathbf{x}\right\|_2^2\right],
\end{aligned} \tag{79}$$

where $\vartheta_{k_t}=q_{k_t}\frac{\mathbf{a}_{k_t}^H\mathbf{x}_t}{\sqrt{|\mathbf{a}_{k_t}^H\mathbf{x}_t|^2+\mu_t^2}}$. Then, according to definition of $\mathbf{h}_t$, Eq. (79) can be rewritten as

$$\begin{aligned}
&\mathbb{E}_{k_t}\left[d_r^2\left(\mathbf{x}_{t+1},\mathbf{x}\right)\right]=\mathbb{E}_{k_t}\left[\left\|\mathbf{h}_t-e^{-j\theta_t}\alpha\left(\mathbf{a}_{k_t}^H\mathbf{x}_t-\vartheta_{k_t}\right)\mathbf{a}_{k_t}\right\|_2^2\right] \\
&=\|\mathbf{h}_t\|_2^2-2\alpha\mathbb{E}_{k_t}\left[\mathscr{R}\left(e^{-j\theta_t}(\mathbf{h}_t^H\mathbf{a}_{k_t})\left(\mathbf{a}_{k_t}^H\mathbf{x}_t-\vartheta_{k_t}\right)\right)\right] \\
&+\alpha^2\mathbb{E}_{k_t}\left[\|\mathbf{a}_{k_t}\|_2^2\left|\mathbf{a}_{k_t}^H\mathbf{x}_t-\vartheta_{k_t}\right|^2\right],
\end{aligned} \tag{80}$$

where $\mathscr{R}(\cdot)$ is the real part function. Since k_t is sampled uniformly at random from $\{1,2,\cdots,m\}$, we have

$$\begin{aligned}
&\mathbb{E}_{k_t}\left[d_r^2\left(\mathbf{x}_t,\mathbf{x}\right)\right] \\
&=\|\mathbf{h}_t\|_2^2+\overbrace{\frac{\alpha^2}{m}\sum_{k=1}^{m}\left[\|\mathbf{a}_k\|_2^2\left|\mathbf{a}_k^H\mathbf{x}_t-q_k\frac{\mathbf{a}_k^H\mathbf{x}_t}{\sqrt{|\mathbf{a}_k^H\mathbf{x}_t|^2+\mu_t^2}}\right|^2\right]}^{v_2} \\
&\underbrace{-\frac{2\alpha}{m}\sum_{k=1}^{m}\mathscr{R}\left(\mathbf{h}_t^H\mathbf{a}_k\left(e^{-j\theta_t}(\mathbf{a}_k^H\mathbf{x}_t)-q_k\frac{e^{-j\theta_t}(\mathbf{a}_k^H\mathbf{x}_t)}{\sqrt{|\mathbf{a}_k^H\mathbf{x}_t|^2+\mu_t^2}}\right)\right)}_{v_1}.
\end{aligned} \tag{81}$$

Notice that, from Eq. (81), v_1 can be rewritten as

$$v_1 = \frac{2\alpha}{m}\sum_{k=1}^{m}|\mathbf{a}_k^H\mathbf{h}_t|^2 - \underbrace{\frac{2\alpha}{m}\sum_{k=1}^{m}\mathscr{R}\left[q_k(\mathbf{h}_t^H\mathbf{a}_k)\left(\frac{e^{-j\theta_t}(\mathbf{a}_k^H\mathbf{x}_t)}{\sqrt{|\mathbf{a}_k^H\mathbf{x}_t|^2+\mu_t^2}} - \frac{\mathbf{a}_k^H\mathbf{x}}{|\mathbf{a}_k^H\mathbf{x}|}\right)\right]}_{h_1}. \tag{82}$$

Given the fact that $\mathscr{R}(w) \leq |w|$ for all $w \in \mathbb{C}$, h_1 in Eq. (82) can be bounded as

$$h_1 \leq \frac{2\alpha}{m}\sum_{k=1}^{m} q_k|\mathbf{a}_k^H\mathbf{h}_t|\left|\frac{e^{-j\theta_t}(\mathbf{a}_k^H\mathbf{x}_t)}{\sqrt{|\mathbf{a}_k^H\mathbf{x}_t|^2+\mu_t^2}} - \frac{\mathbf{a}_k^H\mathbf{x}}{|\mathbf{a}_k^H\mathbf{x}|}\right|. \tag{83}$$

Now, from Eq. (83) it can be obtained that

$$\begin{aligned} h_1 &\leq \frac{2\alpha}{m}\sum_{k=1}^{m} q_k|\mathbf{a}_k^H\mathbf{h}_t|\left|\frac{e^{-j\theta_t}(\mathbf{a}_k^H\mathbf{x}_t)}{\sqrt{|\mathbf{a}_k^H\mathbf{x}_t|^2+\mu_t^2}} - \frac{e^{-j\theta_t}(\mathbf{a}_k^H\mathbf{x}_t)}{|\mathbf{a}_k^H\mathbf{x}|}\right| \\ &+ \frac{2\alpha}{m}\sum_{k=1}^{m} q_k|\mathbf{a}_k^H\mathbf{h}_t|\left|\frac{e^{-j\theta_t}(\mathbf{a}_k^H\mathbf{x}_t)}{|\mathbf{a}_k^H\mathbf{x}|} - \frac{\mathbf{a}_k^H\mathbf{x}}{|\mathbf{a}_k^H\mathbf{x}|}\right| \\ &\leq \frac{2\alpha}{m}\sum_{k=1}^{m}|\mathbf{a}_k^H\mathbf{h}_t|\left|\sqrt{|\mathbf{a}_k^H\mathbf{x}_t|^2+\mu_t^2} - |\mathbf{a}_k^H\mathbf{x}|\right| \\ &+ \frac{2\alpha}{m}\sum_{k=1}^{m}|\mathbf{a}_k^H\mathbf{h}_t|^2, \end{aligned} \tag{84}$$

in which the second inequality comes from the fact that

$$\begin{aligned} & q_k \left| \frac{e^{-j\theta_t}(\mathbf{a}_k^H \mathbf{x}_t)}{\sqrt{|\mathbf{a}_k^H \mathbf{x}_t|^2 + \mu_t^2}} - \frac{e^{-j\theta_t}(\mathbf{a}_k^H \mathbf{x}_t)}{|\mathbf{a}_k^H \mathbf{x}|} \right| \\ & \leq \frac{q_k |\mathbf{a}_k^H \mathbf{x}_t|}{|\mathbf{a}_k^H \mathbf{x}| \sqrt{|\mathbf{a}_k^H \mathbf{x}_t|^2 + \mu_t^2}} \left| \sqrt{|\mathbf{a}_k^H \mathbf{x}_t|^2 + \mu_t^2} - |\mathbf{a}_k^H \mathbf{x}| \right| \\ & \leq \left| \sqrt{|\mathbf{a}_k^H \mathbf{x}_t|^2 + \mu_t^2} - |\mathbf{a}_k^H \mathbf{x}| \right|. \end{aligned} \tag{85}$$

Then, from Eq. (201) it can be obtained that

$$\begin{aligned} \left| \sqrt{|\mathbf{a}_k^H \mathbf{x}_t|^2 + \mu_t^2} - |\mathbf{a}_k^H \mathbf{x}| \right| & \leq \mu_t + ||\mathbf{a}_k^H \mathbf{x}_t| - |\mathbf{a}_k^H \mathbf{x}|| \\ & \leq \mu_0 + |e^{-j\theta_t}(\mathbf{a}_k^H \mathbf{x}_t) - \mathbf{a}_k^H \mathbf{x}| \\ & = \mu_0 + |\mathbf{a}_k^H \mathbf{h}_t|, \end{aligned} \tag{86}$$

in which the second line comes after the triangular inequality. Combining Eqs. (84) and (202) it can be obtained that

$$h_1 \leq \frac{4\alpha}{m} \sum_{k=1}^{m} |\mathbf{a}_k^H \mathbf{h}_t|^2 + \mu_0 \left(\frac{2\alpha}{m} \sum_{k=1}^{m} |\mathbf{a}_k^H \mathbf{h}_t| \right). \tag{87}$$

Putting together Eqs. (81), (82) and (203), one can write that

$$-v_1 \leq -\frac{2\alpha}{m} \sum_{k=1}^{m} |\mathbf{a}_k^H \mathbf{h}_t|^2 + \frac{4\alpha}{m} \sum_{k=1}^{m} |\mathbf{a}_k^H \mathbf{h}_t|^2 + \mu_0 \left(\frac{2\alpha}{m} \sum_{k=1}^{m} |\mathbf{a}_k^H \mathbf{h}_t| \right). \tag{88}$$

On the other hand, since for the $i.i.d.$ real-valued Gaussian vectors $\mathbf{a}_k$, $\max_{k\in\{1,\cdots,m\}}\|\mathbf{a}_k\|^2 \leq 2{,}3n$ holds with probability at least $1-me^{-n/2}$ Wang et al. (2016a), then, from the term v_2 in Eq. (81), we have with high probability that

$$\begin{aligned}
v_2 &\leq \frac{2{,}3n\alpha^2}{m}\sum_{k=1}^{m}\left|\mathbf{a}_k^H\mathbf{x}_t - q_k\frac{\mathbf{a}_k^H\mathbf{x}_t}{\sqrt{|\mathbf{a}_k^H\mathbf{x}_t|^2+\mu_t^2}}\right|^2 \\
&\leq \frac{2{,}3n\alpha^2}{m}\sum_{k=1}^{m}\left|\sqrt{|\mathbf{a}_k^H\mathbf{x}_t|^2+\mu_t^2}-|\mathbf{a}_k^H\mathbf{x}|\right|^2 \\
&\leq \frac{2{,}3n\alpha^2}{m}\sum_{k=1}^{m}\left(\mu_0+|\mathbf{a}_k^H\mathbf{h}_t|\right)^2 \\
&\leq \frac{2{,}3n\alpha^2}{m}\sum_{k=1}^{m}\mu_0^2+\mu_0\left(\frac{4{,}6n\alpha^2}{m}\sum_{k=1}^{m}|\mathbf{a}_k^H\mathbf{h}_t|\right) \\
&+\frac{2{,}3n\alpha^2}{m}\sum_{k=1}^{m}|\mathbf{a}_k^H\mathbf{h}_t|^2,
\end{aligned} \tag{89}$$

where the third inequality was obtained from Eq. (202). Therefore, applying Lemma 7.8 in Candes et al. (2015d), if $m \geq c_0\varepsilon^{-2}n$, then with probability $1-2e^{-\varepsilon^2 m/2}$

$$(1-\varepsilon)\|\mathbf{h}_t\|_2^2 \leq \frac{1}{m}\sum_{k=1}^{m}|\mathbf{a}_k^H\mathbf{h}_t|^2 \leq (1+\varepsilon)\|\mathbf{h}_t\|_2^2, \tag{90}$$

holds for all vectors $\mathbf{h}_t$ and for any $\varepsilon\in(0,1)$. Then, by combining Eqs. (204) and (205) it can be obtained that

$$-v_1 \leq 2\alpha(1+3\varepsilon)\|\mathbf{h}_t\|_2^2+\mu_0\left(\frac{2\alpha}{m}\sum_{k=1}^{m}|\mathbf{a}_k^H\mathbf{h}_t|\right), \tag{91}$$

with probability at least $1-2e^{-\varepsilon^2 m/2}$. Moreover, from the results in Eqs. (192) and (205), it can be concluded that

$$v_2 \leq 2{,}3\alpha^2 n(1+\varepsilon)\|\mathbf{h}_t\|_2^2 + 2{,}3n\alpha^2\mu_0^2 + \mu_0\left(\frac{4{,}6n\alpha^2}{m}\sum_{k=1}^{m}|\mathbf{a}_k^H\mathbf{h}_t|\right). \tag{92}$$

Therefore, by combining Eqs. (81), (206) and (92) we have that

$$\mathbb{E}_{k_t}\left[d_r^2(\mathbf{x}_{t+1},\mathbf{x})\right] \leq \lambda\|\mathbf{h}_t\|_2^2 + \mu_0 c, \tag{93}$$

where $\lambda = 2\alpha(1+3\varepsilon) + 2{,}3n\alpha^2(1+\varepsilon)$ and $c = 2{,}3n\alpha^2\mu_0 + \frac{4{,}6n\alpha}{m}\sum_{k=1}^{m}|\mathbf{a}_k^H\mathbf{h}_t| + \frac{2\alpha}{m}\sum_{k=1}^{m}|\mathbf{a}_k^H\mathbf{h}_t|$. Note that the inequality in Eq. (93) is satisfied for all initial $\mu_0 \in \mathbb{R}_{++}$. Then by Theorem 1.1 in Apostol (1974), one can conclude that

$$\mathbb{E}_{k_t}\left[d_r^2(\mathbf{x}_{t+1},\mathbf{x})\right] \leq \lambda\|\mathbf{h}_t\|_2^2, \tag{94}$$

with probability $1-2e^{-\varepsilon^2 m/2}$. Notice that, choosing the step size $\alpha < \frac{1}{2n(1+3\varepsilon)} = \frac{0{,}5}{n(1+3\varepsilon)} = \frac{\alpha_0}{n}$, then we have that $\lambda \in (0,1)$. Taking $\upsilon = 1-\lambda$, Eq. (94) can be rewritten as

$$\mathbb{E}_{k_t}\left[d_r^2(\mathbf{x}_{t+1},\mathbf{x})\right] \leq (1-\upsilon)\|\mathbf{h}_t\|_2^2. \tag{95}$$

Therefore, from Eq. (95) the lemma is proved. □

Appendix B: Proof of Theorem 4

To prove Theorem 4, this thesis needs to introduce first the contraction mapping definition and the Hahn Banach Fixed Point theorem as follows.

Definition 6. *Contraction mapping:* Let $f : (\mathbb{C}^n, d_r(\cdot,\cdot)) \to \mathbb{R}$ be a function. Then, $f(\mathbf{x})$ is a contraction mapping if there is some non-negative $\beta \in [0,1)$ such that

$$d_r(f(\mathbf{x}), f(\mathbf{y})) \leq \beta d_r(\mathbf{x},\mathbf{y}), \forall \mathbf{x},\mathbf{y} \in \mathbb{R}^n. \tag{96}$$

Theorem 7. *Hahn Banach Fixed Point:* Let $f : (\mathbb{C}^n, d_r(\cdot,\cdot)) \to \mathbb{R}$ be a contraction mapping. Then $f(\mathbf{x})$ admits a unique fixed-point $\mathbf{x}^* \in \mathbb{C}^n$, (*i.e.* $f(\mathbf{x}^*) = \mathbf{x}^*$). (The proof of Theorem 7 can be found in Kreyszig (1989))

Demostración. The sketch of the proof is similar to the used in Appendix C to prove Theorem 7. Denote $\mathscr{K}_1 := \{k | \mu_{k+1} = \gamma_1 \mu_k\}$ with $\gamma_1 \in (0,1)$. If $\mathscr{K}_1$ is finite, then according to Lines 5-8 in Algorithm 2 there exists an integer $\bar{k}$ such that for all $i > \bar{k}$

$$\|\partial g_1(\mathbf{x}_i, \mu_{i-1})\|_2 \geq \gamma \mu_{i-1}, \tag{97}$$

where $\mu_i = \mu_{\bar{k}}$ and $\gamma \in (0,1)$. Taking $\overline{\mu} = \mu_{\bar{k}}$, the optimization problem solved by Algorithm 2,

reduces to solve

$$\min_{\mathbf{x}\in\mathbb{R}^n/\mathbb{C}^n} g_1(\mathbf{x},\overline{\mu}) = \mathbb{E}\left[\ell_{k_t}(\mathbf{x},\overline{\mu})\right]. \tag{98}$$

Assuming the setup of Theorem 3, it can be obtained that Line 4 in Algorithm 2 is contractive according to Definition 6, which means from Theorem 7, that there exists a unique fixed point $\mathbf{x}^*$, up to a global unimodular constant. Then, from Line 4 in Algorithm 2 we have that

$$\mathbf{x}^* = \mathbb{E}\left[\mathbf{x}^* - \alpha\partial\ell_{k_t}(\mathbf{x}^*,\overline{\mu})\right]. \tag{99}$$

Using the fact that $\partial g_1(\mathbf{x}_i,\mu_{i-1}) = \mathbb{E}\left[\partial\ell_{k_t}(\mathbf{x}^*,\overline{\mu})\right]$, from Eq. (99) it can be concluded that

$$\liminf_{i\to\infty}\|\partial g_1(\mathbf{x}_i,\mu_{i-1})\|_2 = 0. \tag{100}$$

Notice that, Eq. (100) contradicts the fact that $\|\partial g_1(\mathbf{x}_i,\mu_{i-1})\|_2 \geq \gamma\mu_{i-1}$ for all $i > \overline{k}$. This shows that $\mathscr{K}_1$ must be infinite and $\lim_{i\to\infty}\mu_i = 0$. Given that $\mathscr{K}_1$ is infinite, it can assume that $\mathscr{K}_1 = \{k_0,k_1,\cdots\}$ with $k_0 < k_1 < \cdots$. Thus, we have that

$$\liminf_{i\to\infty}\|\partial g_1(\mathbf{x}_i,\mu_{i-1})\|_2 \leq \gamma\lim_{i\to\infty}\mu_i = 0. \tag{101}$$

Therefore, from Eq. (101) the result holds. □

Appendix C: Proof of Theorem 5

Demostración. Let $\mathbf{h}^{(t)} = \mathbf{x} - e^{-j\theta_{(t)}}\mathbf{z}^{(t)}$ with $\mathbf{z}^{(t)}$ and $\theta_{(t)} = \arg\min_{\theta\in[0,2\pi)} \|\mathbf{x} - e^{-j\theta}\mathbf{z}^{(t)}\|_2$. Also, define

$$\begin{aligned}\mathbf{d}^{(t)} &= \mathbf{z}^{(t)} - \tau\partial g(\mathbf{z}^{(t)}, \mu_{(t)}) \\ &= \mathbf{z}^{(t)} - \frac{2\tau}{m}\sum_{i=1}^{m}\left(\sqrt{|\mathbf{a}_i^H\mathbf{z}^{(t)}|^2 + \mu_{(t)}^2} - q_i\right)\frac{\mathbf{a}_i^H\mathbf{z}^{(t)}}{\sqrt{|\mathbf{a}_i^H\mathbf{z}^{(t)}|^2 + \mu_{(t)}^2}}\mathbf{a}_i,\end{aligned} \tag{186}$$

for $t = 0, 1, \ldots, \infty$, which stands for the prior estimate to the hard thresholding operation in Algorithm 3, line 4. Let $\Theta_{(t+1)} = S_{(t+1)} \cup S^*$ be a set where $S_{(t+1)}$ is the support of $\mathbf{z}^{(t+1)}$, and S^* is the support of the real solution $\mathbf{x}$. The reconstruction error $\mathbf{h}^{(t+1)}$ is supported on the set $\Theta_{(t+1)} := S^* \cup S_{(t+1)}$; likewise, $\mathbf{h}^{(t)}$ is supported on $\Theta_{(t)} := S^* \cup S_{(t)}$. Moreover, the difference between $\Theta_{(t)}$ and $\Theta_{(t+1)}$ can be defined as $\Theta_{(t)} \setminus \Theta_{(t+1)}$, which consists of all elements of $\Theta_{(t)}$ that are not elements of $\Theta_{(t+1)}$. It is then clear that $|S^*| = |S_{(t)}| = k$, $|\Theta_{(t)}| \leq 2k$, and $|\Theta_{(t)} \setminus \Theta_{(t+1)}| \leq 2k$ as well as $|\Theta_{(t)} \cup \Theta_{(t+1)}| \leq 3k$ for all $t \geq 0$. When using these sets as subscript, for instance, $\mathbf{d}^{(t)}_{\Theta_{(t)}}$, it means vectors formed by setting to zero all but those elements from the vector other than those in the set.

Note that, by definition of $d_r(\cdot,\cdot)$ we have that

$$d_r(\mathbf{z}^{(t+1)}_{\Theta_{(t+1)}}, \mathbf{x}_{\Theta_{(t+1)}}) = \min_{\theta\in[0,2\pi)} \|\mathbf{x}_{\Theta_{(t+1)}} - e^{-j\theta}\mathbf{z}^{(t+1)}_{\Theta_{(t+1)}}\|_2 \leq \|\mathbf{x}_{\Theta_{(t+1)}} - e^{-j\theta_{(t)}}\mathbf{z}^{(t+1)}_{\Theta_{(t+1)}}\|_2. \tag{187}$$

Then, notice that by using the triangle inequality, one can write that

$$\begin{aligned}\|\mathbf{x}_{\Theta_{(t+1)}} - e^{-j\theta_{(t)}}\mathbf{z}^{(t+1)}_{\Theta_{(t+1)}}\|_2 &= \|\mathbf{x}_{\Theta_{(t+1)}} - e^{-j\theta_{(t)}}\mathbf{d}^{(t+1)}_{\Theta_{(t+1)}} + e^{-j\theta_{(t)}}\mathbf{d}^{(t+1)}_{\Theta_{(t+1)}} - e^{-j\theta_{(t)}}\mathbf{z}^{(t+1)}_{\Theta_{(t+1)}}\|_2 \\ &\leq \|\mathbf{x}_{\Theta_{(t+1)}} - e^{-j\theta_{(t)}}\mathbf{d}^{(t+1)}_{\Theta_{(t+1)}}\|_2 \\ &+ \|e^{-j\theta_{(t)}}\mathbf{z}^{(t+1)}_{\Theta_{(t+1)}} - e^{-j\theta_{(t)}}\mathbf{d}^{(t+1)}_{\Theta_{(t+1)}}\|_2, \end{aligned} \tag{188}$$

where in the last inequality the first term is the distance of $\mathbf{x}_{\Theta_{(t+1)}}$ to the estimate $\mathbf{d}^{(t+1)}_{\Theta_{(t+1)}}$ before hard thresholding, and the second is the distance between $\mathbf{d}^{(t+1)}_{\Theta_{(t+1)}}$ and its best k-approximation $\mathbf{z}^{(t+1)}_{\Theta_{(t+1)}}$ due to $|\Theta_{(t+1)}| \leq 2k$. The optimal of $\mathbf{z}^{(t+1)}_{\Theta_{(t+1)}}$ implies $\|e^{-j\theta_{(t)}}\mathbf{z}^{(t+1)}_{\Theta_{(t+1)}} - e^{-j\theta_{(t)}}\mathbf{d}^{(t+1)}_{\Theta_{(t+1)}}\|_2 \leq \|\mathbf{x}_{\Theta_{(t+1)}} - e^{-j\theta_{(t)}}\mathbf{d}^{(t+1)}_{\Theta_{(t+1)}}\|_2$.

Plugging the latter relationship into (188) yields

$$\|\mathbf{x}_{\Theta_{(t+1)}} - e^{-j\theta_{(t)}}\mathbf{z}^{(t+1)}_{\Theta_{(t+1)}}\|_2 \leq 2\|\mathbf{x}_{\Theta_{(t+1)}} - e^{-j\theta_{(t)}}\mathbf{d}^{(t+1)}_{\Theta_{(t+1)}}\|_2, \tag{189}$$

where the equality in (188) arises from restricting this analysis solely to the support $\Theta_{(t+1)}$ of $\mathbf{x} - e^{-j\theta_{(t)}}\mathbf{d}^{(t+1)}$. Then, considering (186), the vector $e^{-j\theta_{(t)}}\mathbf{d}^{(t)}$ can be rewritten as

$$e^{-j\theta_{(t)}}\mathbf{d}^{(t+1)} = e^{-j\theta_{(t)}}\mathbf{z}^{(t)} + \frac{2\tau}{m}\sum_{i=1}^{m}\left(\mathbf{a}_i^H\mathbf{h}^{(t)} + q_i\left(\frac{e^{-j\theta_{(t)}}\mathbf{a}_i^H\mathbf{z}^{(t)}}{\sqrt{|\mathbf{a}_i^H\mathbf{z}^{(t)}|^2 + \mu^2_{(t)}}} - \frac{\mathbf{a}_i^H\mathbf{x}}{|\mathbf{a}_i^H\mathbf{x}|}\right)\right)\mathbf{a}_i. \tag{190}$$

Combining (189) and (190) it can be obtained that

$$\begin{aligned}\frac{1}{2}\|\mathbf{h}^{(t+1)}\|_2 \leq & \Big\|\mathbf{x}_{\Theta^{(t+1)}} - e^{-j\theta_{(t)}}\mathbf{z}^{(t)}_{\Theta^{(t+1)}} - \frac{2\tau}{m}\sum_{i=1}^{m}\left(\mathbf{a}_i^H\mathbf{h}^{(t)}\right)\mathbf{a}_{i,\Theta(t+1)} \\ & -\frac{2\tau}{m}\sum_{i=1}^{m}\left(\frac{e^{-j\theta_{(t)}}\mathbf{a}_i^H\mathbf{z}^{(t)}}{\sqrt{|\mathbf{a}_i^H\mathbf{z}^{(t)}|^2+\mu_{(t)}^2}} - \frac{\mathbf{a}_i^H\mathbf{x}}{|\mathbf{a}_i^H\mathbf{x}|}\right)|\mathbf{a}_i^H\mathbf{x}|\mathbf{a}_{i,\Theta^{(t+1)}}\Big\|_2 \\ = & \Big\|\mathbf{h}^{(t)}_{\Theta^{(t+1)}} - \frac{2\tau}{m}\sum_{i=1}^{m}\mathbf{a}_{i,\Theta^{(t+1)}}\mathbf{a}^H_{i,\Theta^{(t+1)}}\mathbf{h}^{(t)}_{\Theta^{(t+1)}} \\ & -\frac{2\tau}{m}\sum_{i=1}^{m}\mathbf{a}_{i,\Theta(t+1)}\mathbf{a}^H_{i,\Theta^{(t)}\setminus\Theta^{(t+1)}}\mathbf{h}^{(t)}_{\Theta^{(t)}\setminus\Theta^{(t+1)}} \\ & -\frac{2\tau}{m}\sum_{i=1}^{m}\left(\frac{e^{-j\theta_{(t)}}\mathbf{a}_i^H\mathbf{z}^{(t)}}{\sqrt{|\mathbf{a}_i^H\mathbf{z}^{(t)}|^2+\mu_{(t)}^2}} - \frac{\mathbf{a}_i^H\mathbf{x}}{|\mathbf{a}_i^H\mathbf{x}|}\right)|\mathbf{a}_i^H\mathbf{x}|\mathbf{a}_{i,\Theta^{(t+1)}}\Big\|_2,\end{aligned} \tag{191}$$

where the equality follows from re-writing $\mathbf{a}_i^H\mathbf{h}^{(t)} = \mathbf{a}^H_{i,\Theta^{(t)}}\mathbf{h}^{(t)}_{\Theta^{(t)}} = \mathbf{a}^H_{i,\Theta^{(t+1)}}\mathbf{h}^{(t)}_{\Theta^{(t+1)}} + \mathbf{a}^H_{i,\Theta^{(t)}\setminus\Theta^{(t+1)}}\mathbf{h}^{(t)}_{\Theta^{(t)}\setminus\Theta^{(t+1)}}$.

Then, from (191) we have that

$$\begin{aligned}\frac{1}{2}\|\mathbf{h}^{(t+1)}\|_2 \leq & \overbrace{\left\|\mathbf{h}^{(t)}_{\Theta^{(t+1)}} - \frac{2\tau}{m}\sum_{i=1}^{m}\mathbf{a}_{i,\Theta^{(t+1)}}\mathbf{a}^H_{i,\Theta^{(t+1)}}\mathbf{h}^{(t)}_{\Theta^{(t+1)}}\right\|_2}^{v_1} \\ & \underbrace{+\left\|\frac{2\tau}{m}\sum_{i=1}^{m}\mathbf{a}_{i,\Theta^{(t+1)}}\mathbf{a}^H_{i,\Theta^{(t)}\setminus\Theta^{(t+1)}}\mathbf{h}^{(t)}_{\Theta^{(t)}\setminus\Theta^{(t+1)}}\right\|_2}_{v_2} \\ & \underbrace{+\left\|\frac{2\tau}{m}\sum_{i=1}^{m}\left(\frac{e^{-j\theta_{(t)}}\mathbf{a}_i^H\mathbf{z}^{(t)}}{\sqrt{|\mathbf{a}_i^H\mathbf{z}^{(t)}|^2+\mu_{(t)}^2}} - \frac{\mathbf{a}_i^H\mathbf{x}}{|\mathbf{a}_i^H\mathbf{x}|}\right)|\mathbf{a}_i^H\mathbf{x}|\mathbf{a}_{i,\Theta^{(t+1)}}\right\|_2}_{v_3}.\end{aligned} \tag{192}$$

Notice that from (192) it can be obtained that

$$\begin{aligned} v_1 &= \left\| \left(\mathbf{I}_n - \frac{2\tau}{m} \sum_{i=1}^{m} \mathbf{a}_{i,\Theta^{(t+1)}} \mathbf{a}_{i,\Theta^{(t+1)}}^H \right) \mathbf{h}_{\Theta^{(t+1)}}^{(t)} \right\|_2 \\ &\leq \left\| \mathbf{I}_n - \frac{2\tau}{m} \sum_{i=1}^{m} \mathbf{a}_{i,\Theta^{(t+1)}} \mathbf{a}_{i,\Theta^{(t+1)}}^H \right\|_{2\to 2} \|\mathbf{h}_{\Theta^{(t+1)}}^{(t)}\|_2 \\ &\leq \text{máx}\{1 - 2\tau\underline{\lambda}, 2\tau\overline{\lambda} - 1\} \|\mathbf{h}_{\Theta^{(t+1)}}^{(t)}\|_2, \end{aligned} \tag{193}$$

where $\|\cdot\|_{2\to 2}$ is the spectral norm and $\overline{\lambda}, \underline{\lambda} > 0$ are the largest and the smallest eigenvalues of $\frac{1}{m}\sum_{i=1}^{m} \mathbf{a}_{i,\Theta^{(t+1)}} \mathbf{a}_{i,\Theta^{(t+1)}}^H$, respectively. Then, by corollary 5.35 in Vershynin (2010) it can be obtained that

$$\overline{\lambda} = \lambda_{max} \left(\frac{1}{m} \sum_{i=1}^{m} \mathbf{a}_{i,\Theta^{(t+1)}} \mathbf{a}_{i,\Theta^{(t+1)}}^H \right) \leq 1 + \varepsilon_0, \tag{194}$$

with hight probability when $m \geq C(\varepsilon_0)2k$ for some constant $C(\varepsilon_0)$ depending on $\varepsilon_0 > 0$. Moreover, by Lemma 5 in Wang et al. (2016a) we have that

$$\underline{\lambda} = \lambda_{min} \left(\frac{1}{m} \sum_{i=1}^{m} \mathbf{a}_{i,\Theta^{(t+1)}} \mathbf{a}_{i,\Theta^{(t+1)}}^H \right) \geq 1 - \zeta_1 - \varepsilon_1 \tag{195}$$

when $m \geq C(\varepsilon_1)k$ for some constant $C(\varepsilon_1)$ depending on $\varepsilon_1 > 0$. Taking the results in (194) and (195) into (193) yields

$$v_1 \leq \text{máx}\{1 - 2\tau(1 - \zeta_1 - \varepsilon_1), 2\tau(1 + \varepsilon_0) - 1\} \|\mathbf{h}_{\Theta^{(t+1)}}^{(t)}\|_2. \tag{196}$$

For the second term v_2 in (192), fix any $\varepsilon_2 > 0$. If the ratio number of measurements and

unknowns $m/3k$, exceeds some sufficiently large constant, the next holds with probability of at least $1-2\exp(-c(\varepsilon_2)m)$

$$\begin{aligned} v_2 \leq & \left\| \frac{2\tau}{m} \sum_{i=1}^{m} \mathbf{a}_{i,\Theta^{(t+1)}} \mathbf{a}_{i,\Theta^{(t)} \setminus \Theta^{(t+1)}}^H \right\|_{2\to 2} \| \mathbf{h}_{\Theta^{(t)} \setminus \Theta^{(t+1)}}^{(t)} \|_2 \\ \leq & 2\tau \left\| \mathbf{I}_n - \frac{1}{m} \sum_{i=1}^{m} \mathbf{a}_{i,\Theta^{(t+1)} \cup \Theta^{(t)}} \mathbf{a}_{i,\Theta^{(t+1)} \cup \Theta^{(t)}}^H \right\|_{2\to 2} \| \mathbf{h}_{\Theta^{(t)} \setminus \Theta^{(t+1)}}^{(t)} \|_2 \\ \leq & 2\tau(\zeta_2 + \varepsilon_2) \| \mathbf{h}_{\Theta^{(t)} \setminus \Theta^{(t+1)}}^{(t)} \|_2, \end{aligned} \tag{197}$$

in which the first inequality arises from the triangle inequality. The second inequality is obtained by Lemma 1 in Blumensath and Davies (2009). Similar to (193), the last inequality in (197) is obtained by using corollary 5.35 in Vershynin (2010) for some universal constants $c(\varepsilon_2)$ and $C(\varepsilon_2)$ such as $m \geq C(\varepsilon_2)2k$.

Considering the last term v_3 in (192), define $\mathbf{A} := [\mathbf{a}_{1,\Theta^{(t+1)}}, \cdots, \mathbf{a}_{m,\Theta^{(t+1)}}]$ and $\mathbf{b}^{(t)} := [b_1^{(t)}, \cdots, b_m^{(t)}]^T$ with $b_i^{(t)} = \left(\frac{e^{-j\theta_{(t)}} \mathbf{a}_i^H \mathbf{z}^{(t)}}{\sqrt{|\mathbf{a}_i^H \mathbf{z}^{(t)}|^2 + \mu_{(t)}^2}} - \frac{\mathbf{a}_i^H \mathbf{x}}{|\mathbf{a}_i^H \mathbf{x}|} \right) |\mathbf{a}_i^H \mathbf{x}|$, for $i = 1, \cdots, m$. Then, the v_3 term in (192) can be rewritten as

$$\begin{aligned} v_3 = \left\| \frac{2\tau}{m} \mathbf{A}_{\Theta^{(t+1)}}^T \mathbf{b}^{(t)} \right\|_2 \leq & 2\tau \left\| \frac{1}{\sqrt{m}} \mathbf{A}_{\Theta^{(t+1)}}^T \right\|_{2\to 2} \left\| \frac{1}{\sqrt{m}} \mathbf{b}^{(t)} \right\|_2 \\ \leq & 2\tau(1+\varepsilon_3) \left\| \frac{1}{\sqrt{m}} \mathbf{b}^{(t)} \right\|_2, \end{aligned} \tag{198}$$

where the second inequality is obtained by a standard matrix concentration result for any fixed $\varepsilon_3 > 0$, with probability $1-2\exp(-c(\varepsilon_3)m)$, provided that $m \geq C(\varepsilon_3)k$, for some sufficiently large constant $C(\varepsilon_3) > 0$.

Notice that, from the definition of vector $\mathbf{b}^{(t)}$ it can be obtained that

$$\frac{1}{m}\left\|\mathbf{b}^{(t)}\right\|_2^2 = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{e^{-j\theta_{(t)}}\mathbf{a}_i^H\mathbf{z}^{(t)}}{\sqrt{|\mathbf{a}_i^H\mathbf{z}^{(t)}|^2+\mu_{(t)}^2}} - \frac{\mathbf{a}_i^H\mathbf{x}}{|\mathbf{a}_i^H\mathbf{x}|}\right|^2 |\mathbf{a}_i^H\mathbf{x}|^2. \tag{199}$$

Notice that from (199) one can write that

$$\begin{aligned}\left|\frac{e^{-j\theta_{(t)}}\mathbf{a}_i^H\mathbf{z}^{(t)}}{\sqrt{|\mathbf{a}_i^H\mathbf{z}^{(t)}|^2+\mu_{(t)}^2}} - \frac{\mathbf{a}_i^H\mathbf{x}}{|\mathbf{a}_i^H\mathbf{x}|}\right||\mathbf{a}_i^H\mathbf{x}| &\leq |\mathbf{a}_i^H\mathbf{x}|\left|\frac{e^{-j\theta_{(t)}}\mathbf{a}_i^H\mathbf{z}^{(t)}}{\sqrt{|\mathbf{a}_i^H\mathbf{z}^{(t)}|^2+\mu_{(t)}^2}} - \frac{e^{-j\theta_{(t)}}\mathbf{a}_i^H\mathbf{z}^{(t)}}{|\mathbf{a}_i^H\mathbf{x}|}\right| \\ &\quad + |\mathbf{a}_i^H\mathbf{x}|\left|\frac{e^{-j\theta_{(t)}}\mathbf{a}_i^H\mathbf{z}^{(t)}}{|\mathbf{a}_i^H\mathbf{x}|} - \frac{\mathbf{a}_i^H\mathbf{x}}{|\mathbf{a}_i^H\mathbf{x}|}\right| \\ &\leq \left|\sqrt{|\mathbf{a}_i^H\mathbf{z}^{(t)}|^2+\mu_{(t)}^2} - |\mathbf{a}_i^H\mathbf{x}|\right| + |\mathbf{a}_i^H\mathbf{h}^{(t)}|\end{aligned} \tag{200}$$

in which the second inequality comes from the fact that

$$\begin{aligned}|\mathbf{a}_i^H\mathbf{x}|\left|\frac{e^{-j\theta_{(t)}}\mathbf{a}_i^H\mathbf{z}^{(t)}}{\sqrt{|\mathbf{a}_i^H\mathbf{z}^{(t)}|^2+\mu_{(t)}^2}} - \frac{e^{-j\theta_{(t)}}\mathbf{a}_i^H\mathbf{z}^{(t)}}{|\mathbf{a}_i^H\mathbf{x}|}\right| &\leq \frac{|\mathbf{a}_i^H\mathbf{x}||\mathbf{a}_i^H\mathbf{z}^{(t)}|}{|\mathbf{a}_i^H\mathbf{x}|\sqrt{|\mathbf{a}_i^H\mathbf{z}^{(t)}|^2+\mu_{(t)}^2}}\left|\sqrt{|\mathbf{a}_i^H\mathbf{z}^{(t)}|^2+\mu_{(t)}^2} - |\mathbf{a}_i^H\mathbf{x}|\right| \\ &\leq \left|\sqrt{|\mathbf{a}_i^H\mathbf{z}^{(t)}|^2+\mu_{(t)}^2} - |\mathbf{a}_i^H\mathbf{x}|\right|.\end{aligned} \tag{201}$$

Then, from (201) it can be obtained that

$$\begin{aligned}\left|\sqrt{|\mathbf{a}_i^H\mathbf{z}^{(t)}|^2+\mu_{(t)}^2} - |\mathbf{a}_i^H\mathbf{x}|\right| &\leq \mu_{(t)} + \left||\mathbf{a}_i^H\mathbf{z}^{(t)}| - |\mathbf{a}_i^H\mathbf{x}|\right| \\ &\leq \mu_{(0)} + \left|e^{-j\theta_{(t)}}\mathbf{a}_i^H\mathbf{z}^{(t)} - \mathbf{a}_i^H\mathbf{x}\right| \\ &= \mu_{(0)} + |\mathbf{a}_i^H\mathbf{h}^{(t)}|,\end{aligned} \tag{202}$$

in which the second line comes after the triangular inequality. Then, putting together (200) and (202) one can conclude that

$$\left| \frac{e^{-j\theta_{(t)}}\mathbf{a}_i^H\mathbf{z}^{(t)}}{\sqrt{|\mathbf{a}_i^H\mathbf{z}^{(t)}|^2+\mu_{(t)}^2}} - \frac{\mathbf{a}_i^H\mathbf{x}}{|\mathbf{a}_i^H\mathbf{x}|} \right| |\mathbf{a}_i^H\mathbf{x}| \leq \mu_{(0)} + 2|\mathbf{a}_i^H\mathbf{h}^{(t)}|. \tag{203}$$

Combining (199) and (203) it can be obtained that

$$\frac{1}{m}\left\|\mathbf{b}^{(t)}\right\|_2^2 \leq \frac{1}{m}\sum_{i=1}^{m}\left(\mu_{(0)} + 2|\mathbf{a}_i^H\mathbf{h}^{(t)}|\right)^2 = \frac{4}{m}\sum_{i=1}^{m}|\mathbf{a}_i^H\mathbf{h}^{(t)}|^2 + \mu_{(0)}c, \tag{204}$$

where $c = \mu_{(0)}\left(\frac{4}{m}\sum_{i=1}^{m}|\mathbf{a}_i^H\mathbf{h}^{(t)}| + \mu_{(0)}\right)$. Applying Lemma 7.8 in Candes et al. (2015d), if $m \geq c_0\varepsilon_4^{-2}n$, then with probability $1-2e^{-\varepsilon_4^2 m/2}$

$$(1-\varepsilon_4)\|\mathbf{h}^{(t)}\|_2^2 \leq \frac{1}{m}\sum_{i=1}^{m}|\mathbf{a}_i^H\mathbf{h}^{(t)}|^2 \leq (1+\varepsilon_4)\|\mathbf{h}^{(t)}\|_2^2, \tag{205}$$

holds for all vectors $\mathbf{h}^{(t)}$ and for any $\varepsilon_4 \in (0,1)$. Then, by combining (204) and (205) it can be obtained that

$$\frac{1}{m}\left\|\mathbf{b}^{(t)}\right\|_2^2 \leq 4(1+\varepsilon_4)\|\mathbf{h}^{(t)}\|_2^2 + \mu_{(0)}c \tag{206}$$

with probability at least $1-2e^{-\varepsilon_4^2 m/2}$.

Notice that inequality in (206) is satisfied for all initial $\mu_{(0)} \in \mathbb{R}_{++}$. Then, by Theorem 1.1

in Apostol (1974), one can conclude that

$$\begin{aligned} \frac{1}{m}\left\|\mathbf{b}^{(t)}\right\|_2^2 &\leq 4(1+\varepsilon_4)\|\mathbf{h}^{(t)}_{\Theta^{(t)}}\|_2^2 \\ \left\|\frac{1}{\sqrt{m}}\mathbf{b}^{(t)}\right\|_2 &\leq 2(1+\varepsilon_5)\|\mathbf{h}^{(t)}_{\Theta^{(t)}}\|_2, \end{aligned} \tag{207}$$

for any $\varepsilon_5 > 0$ with probability at least $1-2e^{-\varepsilon_5^2 m/2}$.

Therefore, putting together the bounds in (196), (197), (198) and (207) into (192), one can write

$$\begin{aligned} \frac{1}{2}\|\mathbf{h}^{(t+1)}\|_2 &\leq \text{máx}\{1-2\tau(1-\zeta_1-\varepsilon_1), 2\tau(1+\varepsilon_0)-1\}\|\mathbf{h}^{(t)}_{\Theta^{(t+1)}}\|_2 \\ &\quad +2\tau(\zeta_2+\varepsilon_2)\|\mathbf{h}^{(t)}_{\Theta^{(t)}\setminus\Theta^{(t+1)}}\|_2 + 4\tau(1+\varepsilon_3)(1+\varepsilon_5)\|\mathbf{h}^{(t)}_{\Theta^{(t)}}\|_2 \\ &\leq \sqrt{2}\,\text{máx}\{\vartheta, 2\tau(\zeta_2+\varepsilon_2)\}\|\mathbf{h}^{(t)}\|_2 + 4\tau(1+\varepsilon_3)(1+\varepsilon_5)\|\mathbf{h}^{(t)}\|_2 \\ \|\mathbf{h}^{(t+1)}\|_2 &\leq \rho\|\mathbf{h}^{(t)}\|_2, \end{aligned} \tag{208}$$

in which the second inequality results from $\|\mathbf{h}^{(t)}_{\Theta^{(t+1)}}\|_2 + \|\mathbf{h}^{(t)}_{\Theta^{(t)}\setminus\Theta^{(t+1)}}\|_2 \leq \sqrt{2}\|\mathbf{h}^{(t)}\|_2$, with $\vartheta = \text{máx}\{1-2\tau(1-\zeta_1-\varepsilon_1), 2\tau(1+\varepsilon_0)-1\}$. From the last inequality it can be obtained that

$$\rho = 2\left(\sqrt{2}\,\text{máx}\{\vartheta, 2\tau(\zeta_2+\varepsilon_2)\} + 4\tau(1+\varepsilon_3)(1+\varepsilon_5)\right). \tag{209}$$

Then, to ensure linear convergence, from (209) it suffices to choose a step $\tau > 0$ such that $\rho < 1$ in (209). Letting $\eta = 1-\rho \in (0,1)$, which justifies the linear convergence result in (89) with

probability exceeding $1 - 2e^{-c_1 m}$ for some $c_1 \geq 0$. □

Bibliographic References

Apostol, T. M. (1974). Mathematical analysis.

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.

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

Bacca, J., Correa, C. V., and Arguello, H. (2019). Noniterative hyperspectral image reconstruction from compressive fused measurements. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(4):1231–1239.

Bacca, J., Fonseca, Y., and Arguello, H. (2021). Compressive spectral image reconstruction using deep prior and low-rank tensor representation. *arXiv preprint arXiv:2101.07424*.

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., Pinilla, S., Molina, D., Camacho, A., and Arguello, H. (2018). Super-resolution phase retrieval algorithm using a smoothing function. In *Mathematics in Imaging*, pages MW2D–3. Optical Society of America.

Bacca, J., Vargas, H., and Arguello, H. (2017). A constrained formulation for compressive spectral image reconstruction using linear mixture models. In *Proc. Conf. CAMSAP*, pages 1–5. IEEE.

Baraniuk, R. G. (2007). Compressive sensing. *IEEE signal processing magazine*, 24(4).

Baraniuk, R. G., Goldstein, T., Sankaranarayanan, A. C., Studer, C., Veeraraghavan, A., and Wakin, M. B. (2017). Compressive video sensing: algorithms, architectures, and applications. *IEEE Signal Processing Magazine*, 34(1):52–66.

Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

Bioucas-Dias, J. M. and Nascimento, J. M. (2008). Hyperspectral subspace identification. *IEEE Transactions on Geoscience and Remote Sensing*, 46(8):2435–2445.

Bioucas-Dias, J. M., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P., and Chanussot, J. (2012). Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE journal of selected topics in applied earth observations and remote sensing*, 5(2):354–379.

Bioucas-Dias, J. M. and Valadao, G. (2007). Phase Unwrapping via Graph Cuts. *Image Processing, IEEE Transactions on*, 16(3):698–709.

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., Chu, E., Peleato, B., Eckstein, J., et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Buades, A., Coll, B., and Morel, J.-M. (2011). Non-local means denoising. *Image Processing On Line*, 1:208–212.

Candès, E. J. and Li, X. (2014). Solving quadratic equations via phaselift when there are about as many equations as unknowns. *Foundations of Computational Mathematics*, 14(5):1017–1026.

Candes, E. J., Li, X., and Soltanolkotabi, M. (2015a). Phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, 39(2):277–299.

Candes, E. J., Li, X., and Soltanolkotabi, M. (2015b). Phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, 39(2):277–299.

Candes, E. J., Li, X., and Soltanolkotabi, M. (2015c). Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007.

Candes, E. J., Li, X., and Soltanolkotabi, M. (2015d). Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007.

Candes, E. J., Li, X., and Soltanolkotabi, M. (2015e). Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007.

Candes, E. J., Strohmer, T., and Voroninski, V. (2013). Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274.

Candès, E. J. and Wakin, M. B. (2008). An introduction to compressive sampling. *IEEE signal processing magazine*, 25(2):21–30.

Cao, X., Yue, T., Lin, X., Lin, S., Yuan, X., Dai, Q., Carin, L., and Brady, D. J. (2016). Computational snapshot multispectral cameras: Toward dynamic capture of the spectral world. *IEEE Signal Processing Magazine*, 33(5):95–108.

Chakrabarti, A. and Zickler, T. (2011). Statistics of real-world hyperspectral images. In *CVPR 2011*, pages 193–200. IEEE.

Chen, J., Wang, L., Zhang, X., and Gu, Q. (2017). Robust wirtinger flow for phase retrieval with arbitrary corruption. *arXiv preprint arXiv:1704.06256*.

Chen, X. and Zhou, W. (2010). Smoothing nonlinear conjugate gradient method for image restoration using nonsmooth nonconvex minimization. *SIAM Journal on Imaging Sciences*, 3(4):765–790.

Chen, Y. and Candes, E. (2015a). Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems*, pages 739–747.

Chen, Y. and Candes, E. (2015b). Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems*, pages 739–747.

Choi, I., Jeon, D. S., Nam, G., Gutierrez, D., and Kim, M. H. (2017). High-quality hyperspectral reconstruction using a spectral prior. *ACM Transactions on Graphics (TOG)*, 36(6):1–13.

Correa, C. V., Arguello, H., and Arce, G. R. (2015). Snapshot colored compressive spectral imager. *JOSA A*, 32(10):1754–1763.

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

Correa, C. V., Hinojosa, C., Arce, G. R., and Arguello, H. (2016b). Multiple snapshot colored compressive spectral imager. *Optical Engineering*, 56(4):041309.

Diaz, N., Hinojosa, C., and Arguello, H. (2019). Adaptive grayscale compressive spectral imaging using optimal blue noise coding patterns. *Optics & Laser Technology*, 117:147–157.

Donoho, D. L. and Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81(3):425–455.

Duarte, M. F., Davenport, M. A., Takhar, D., Laska, J. N., Sun, T., Kelly, K. F., and Baraniuk, R. G. (2008). Single-pixel imaging via compressive sampling. *IEEE signal processing magazine*, 25(2):83–91.

Duchi, J. C. and Ruan, F. (2017). Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *arXiv preprint arXiv:1705.02356*.

Dürig, U., Pohl, D. W., and Rohner, F. (1986). Near-field optical-scanning microscopy. *Journal of applied physics*, 59(10):3318–3327.

Ekeland, I. and Temam, R. (1999). *Convex analysis and variational problems*. SIAM.

Eldar, Y. C. and Bolcskei, H. (2009). Block-sparsity: Coherence and efficient recovery. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2885–2888. IEEE.

Elhamifar, E. and R.Vidal (2013). Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2765–2781.

Elhamifar, E. and Vidal, R. (2009). Sparse subspace clustering. In *Proc. Conf. CVPR*, pages 2790–2797. IEEE.

Eriksson, K., Estep, D., and Johnson, C. (2013). *Applied mathematics: Body and soul: Volume 1: Derivatives and geometry in IR3*. Springer Science & Business Media.

Estrada, J. C., Servin, M., and Quiroga, J. A. (2011). Noise robust linear dynamic system for phase unwrapping and smoothing. *Optics express*, 19(6):5126–5133.

Fienup, C. and Dainty, J. (1987). Phase retrieval and image reconstruction for astronomy. *Image Recovery: Theory and Application*, pages 231–275.

Fienup, J. R. (1982). Phase retrieval algorithms: a comparison. *Applied optics*, 21(15):2758–2769.

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.

Gao, W., Huyen, N. T. T., Loi, H. S., and Kemao, Q. (2009). Real-time 2D parallel windowed Fourier transform for fringe pattern analysis using Graphics Processing Unit. *Optics express*, 17(25):23147–23152.

Gehm, M., John, R., Brady, D., Willett, R., and Schulz, T. (2007). Single-shot compressive spectral imaging with a dual-disperser architecture. *Optics express*, 15(21):14013–14027.

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

Ghiglia, D. C. and Pritt, M. D. (1998). *Two-dimensional phase unwrapping: theory, algorithms, and software*, volume 4. Wiley New York.

Ghiglia, D. C. and Romero, L. A. (1996). Minimum lp-norm two-dimensional phase unwrapping. *JOSA A*, 13(10):1999–2013.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Goodman, J. (2008). Introduction to fourier optics.

Goodman, J. W. (2005). Introduction to fourier optics. *Introduction to Fourier optics, 3rd ed., by JW Goodman. Englewood, CO: Roberts & Co. Publishers, 2005*, 1.

Griffin, D. and Lim, J. (1984). Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243.

Grippo, L., Lampariello, F., and Lucidi, S. (1986). A nonmonotone line search technique for newton's method. *SIAM Journal on Numerical Analysis*, 23(4):707–716.

Gross, D., Krahmer, F., and Kueng, R. (2017). Improved recovery guarantees for phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, 42(1):37–64.

Gu, D., Gillespie, A. R., Kahle, A. B., and Palluconi, F. D. (2000). Autonomous atmospheric compensation (aac) of high resolution hyperspectral thermal infrared remote-sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 38(6):2557–2570.

He, W., Zhang, H., Zhang, L., and Shen, H. (2015). Hyperspectral image denoising via noise-adjusted iterative low-rank matrix approximation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(6):3050–3061.

Hernandez-Lopez, F. J., Rivera, M., Salazar-Garibay, A., and Legarda-Sáenz, R. (2018). Comparison of multihardware parallel implementations for a phase unwrapping algorithm. *Optical Engineering*, 57(4):043113.

Hess, H., Betzig, E., Harris, T., Pfeiffer, L., and West, K. (1994). Near-field spectroscopy of the quantum constituents of a luminescent system. *Science*, 264(5166):1740–1745.

Higham, C. F., Murray-Smith, R., Padgett, M. J., and Edgar, M. P. (2018). Deep learning for real-time single-pixel video. *Scientific reports*, 8(1):1–9.

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.

Hongxing, H. and Lingda, W. (2014). PUMA-SPA: A Phase Unwrapping Method Based on PUMA and Second-Order Polynomial Approximation. *IEEE Geoscience and Remote Sensing Letters*, 11(11):1906–1910.

Hunger, R. (2007). *An introduction to complex differentials and complex differentiability*. Munich University of Technology, Inst. for Circuit Theory and Signal Processing.

Jaganathan, K., Saunderson, J., Fazei, M., Eldar, Y. C., and Hassibi, B. (2016). Phaseless super-resolution using masks. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 4039–4043. IEEE.

Jahncke, C., Paesler, M., and Hallen, H. (1995). Raman imaging with near-field scanning optical microscopy. *Applied physics letters*, 67(17):2483–2485.

Jerez, A., Pinilla, S., and Arguello, H. (2020). Fast target detection via template matching in compressive phase retrieval. *IEEE Transactions on Computational Imaging*, 6:934–944.

Kaski, S. (1998). Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Proc. Conf. IJCNN*, volume 1, pages 413–418. IEEE.

Katkovnik, V., Shevkunov, I., Petrov, N. V., and Egiazarian, K. (2017). Computational superresolution phase retrieval from multiple phase-coded diffraction patterns: simulation study and experiments. *Optica*, 4(7):786–794.

Katkovnik, V. Y. and Egiazarian, K. (2017). Sparse superresolution phase retrieval from phase-coded noisy intensity patterns. *Optical Engineering*, 56(9):094103.

Kemao, Q. (2004). Windowed fourier transform for fringe pattern analysis. *Applied Optics*, 43(13):2695–2702.

Kemao, Q. (2007). Two-dimensional windowed fourier transform for fringe pattern analysis: principles, applications and implementations. *Optics and Lasers in Engineering*, 45(2):304–317.

Kemao, Q., Gao, W., and Wang, H. (2008). Windowed fourier-filtered and quality-guided phase-unwrapping algorithm. *Applied optics*, 47(29):5420–5428.

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

Kittle, D., Choi, K., Wagadarikar, A., and Brady, D. J. (2010). Multiframe image estimation for coded aperture snapshot spectral imagers. *Applied Optics*, 49(36):6824–6833.

Kreyszig, E. (1989). *Introductory functional analysis with applications*, volume 1. wiley New York.

Ledoux, M. (2005). *The concentration of measure phenomenon*. Number 89. American Mathematical Soc.

León-López, K. M. and Fuentes, H. A. (2020). Online tensor sparsifying transform based on temporal superpixels from compressive spectral video measurements. *IEEE Transactions on Image Processing*, 29:5953–5963.

Li, X. and Voroninski, V. (2013). Sparse signal recovery from quadratic measurements via convex programming. *SIAM Journal on Mathematical Analysis*, 45(5):3019–3033.

Lin, X., Liu, Y., Wu, J., and Dai, Q. (2014a). Spatial-spectral encoded compressive hyperspectral imaging. *ACM Transactions on Graphics (TOG)*, 33(6):233.

Lin, X., Wetzstein, G., Liu, Y., and Dai, Q. (2014b). Dual-coded compressive hyperspectral imaging. *Optics letters*, 39(7):2044–2047.

Loewen, E. G. and Popov, E. (2018). *Diffraction gratings and applications*. CRC Press.

Lohit, S., Singh, R., Kulkarni, K., and Turaga, P. (2019). Rank-regularized measurement operators for compressive imaging. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pages 942–946. IEEE.

Martín, G. and Bioucas-Dias, J. M. (2016). Hyperspectral blind reconstruction from random spectral projections. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(6):2390–2399.

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

Meng, L., Fang, S., Yang, P., Wang, L., Komori, M., and Kubo, A. (2012). Image-inpainting and quality-guided phase unwrapping algorithm. *Applied optics*, 51(13):2457–2462.

Miao, J., Chapman, H., and Sayre, D. (2000). Image reconstruction from the oversampled diffraction pattern. *MICROSCOPY AND MICROANALYSIS-NEW YORK-*, 3(2):1155–1156.

Mojica, E., Pertuz, S., and Arguello, H. (2017). High-resolution coded-aperture design for compressive x-ray tomography using low resolution detectors. *Optics Communications*, 404:103–109.

Montresor, S. and Picart, P. (2016). Quantitative appraisal for noise reduction in digital holographic phase imaging. *Optics express*, 24(13):14322–14343.

Nascimento, J. M. P. and Dias, J. M. B. (2005). Vertex component analysis: a fast algorithm to unmix hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 43(4):898–910.

Netrapalli, P., Jain, P., and Sanghavi, S. (2013). Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems*, pages 2796–2804.

Ng, A. Y., Jordan, M. I., Weiss, Y., et al. (2002). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856.

Ohlsson, H., Yang, A. Y., Dong, R., and Sastry, S. S. (2011). Compressive phase retrieval from squared output measurements via semidefinite programming. *arXiv preprint arXiv:1111.6323*, pages 1–27.

Pineda, J., Bacca, J., Meza, J., Romero, L. A., Arguello, H., and Marrugo, A. G. (2020). Spud: simultaneous phase unwrapping and denoising algorithm for phase imaging. *Applied optics*, 59(13):D81–D88.

Pinilla, S., Bacca, J., and Arguello, H. (2018a). Phase retrieval algorithm via nonconvex minimization using a smoothing function. *IEEE Transactions on Signal Processing*, 66(17):4574–4584.

Pinilla, S., Poveda, J., and Arguello, H. (2018b). Coded diffraction system in x-ray crystallography using a boolean phase coded aperture approximation. *Optics Communications*, 410:707–716.

Poon, T.-C. and Liu, J.-P. (2014). *Introduction to modern digital holography: with MATLAB*. Cambridge University Press.

Qian, C., Fu, X., Sidiropoulos, N. D., Huang, L., and Xie, J. (2017). Inexact alternating optimization for phase retrieval in the presence of outliers. *IEEE Transactions on Signal Processing*, 65(22):6069–6082.

Qian, Y., Jia, S., Zhou, J., and Robles-Kelly, A. (2011). Hyperspectral unmixing via $l\_\{1/2\}$ sparsity-constrained nonnegative matrix factorization. *IEEE Transactions on Geoscience and Remote Sensing*, 49(11):4282–4297.

Rodenburg, J. (2008). Ptychography and related diffractive imaging methods. *Advances in Imaging and Electron Physics*, 150:87–184.

Rueda, H., Arguello, H., and Arce, G. R. (2015a). Dmd-based implementation of patterned optical filter arrays for compressive spectral imaging. *JOSA A*, 32(1):80–89.

Rueda, H., Lau, D., and Arce, G. R. (2015b). Multi-spectral compressive snapshot imaging using rgb image sensors. *Optics express*, 23(9):12207–12221.

SAR-EDU (2019). Sar-edu remote sensing education initiative, dem generation with matlab.

Shechtman, Y., Eldar, Y. C., Cohen, O., Chapman, H. N., Miao, J., and Segev, M. (2015). Phase retrieval with application to optical imaging: a contemporary overview. *IEEE signal processing magazine*, 32(3):87–109.

Shimano, T., Nakamura, Y., Tajima, K., Sao, M., and Hoshizawa, T. (2018). Lensless light-field imaging with fresnel zone aperture: quasi-coherent coding. *Applied optics*, 57(11):2841–2850.

Shuyue, C. and Hongnian, L. (2000). Noise characteristic and its removal in digital radiographic system. In *15th World Conference on Nondestructive Testing*.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer.

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.

Soldevila, F., Irles, E., Durán, V., Clemente, P., Fernández-Alonso, M., Tajahuerce, E., and Lancis, J. (2013). Single-pixel polarimetric imaging spectrometer by compressive sensing. *Applied Physics B*, 113(4):551–558.

Sun, T. and Kelly, K. (2009). Compressive sensing hyperspectral imager. In *Computational Optical Sensing and Imaging*, page CTuA5. Optical Society of America.

Szameit, A., Shechtman, Y., Osherovich, E., Bullkich, E., Sidorenko, P., Dana, H., Steiner, S., Kley, E. B., Gazit, S., Cohen-Hyams, T., et al. (2012). Sparsity-based single-shot subwavelength coherent diffractive imaging. *Nature materials*, 11(5):455.

Thibault, P., Dierolf, M., Bunk, O., Menzel, A., and Pfeiffer, F. (2009). Probe retrieval in ptychographic coherent diffractive imaging. *Ultramicroscopy*, 109(4):338–343.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

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

Wang, G., Giannakis, G. B., and Chen, J. (2017a). Scalable solvers of random quadratic equations via stochastic truncated amplitude flow. *IEEE Transactions on Signal Processing*, 65(8):1961–1974.

Wang, G., Giannakis, G. B., and Eldar, Y. C. (2016a). Solving systems of random quadratic equations via truncated amplitude flow. *arXiv preprint arXiv:1605.08285*.

Wang, G., Giannakis, G. B., and Eldar, Y. C. (2017b). Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Transactions on Information Theory*, 64(2):773–794.

Wang, G., Giannakis, G. B., and Eldar, Y. C. (2018a). Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Transactions on Information Theory*, 64(2):773–794.

Wang, G., Giannakis, G. B., Saad, Y., and Chen, J. (2018b). Phase retrieval via reweighted amplitude flow. *IEEE Transactions on Signal Processing*, 66(11):2818–2833.

Wang, G., Zhang, L., Giannakis, G. B., Akçakaya, M., and Chen, J. (2016b). Sparse phase retrieval via truncated amplitude flow. *arXiv preprint arXiv:1611.07641*.

Wang, G., Zhang, L., Giannakis, G. B., Akçakaya, M., and Chen, J. (2017c). Sparse phase retrieval via truncated amplitude flow. *IEEE Transactions on Signal Processing*, 66(2):479–491.

Wang, L., Feng, Y., Gao, Y., Wang, Z., and He, M. (2018c). Compressed sensing reconstruction of hyperspectral images based on spectral unmixing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(4):1266–1284.

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

Wang, L., Sun, C., Zhang, M., Fu, Y., and Huang, H. (2020). 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. (2018d). Hyperreconnet: Joint coded aperture optimization and image reconstruction for compressive hyperspectral imaging. *IEEE Transactions on Image Processing*, 28(5):2257–2270.

Wang, Y., Lin, L., Zhao, Q., Yue, T., Meng, D., and Leung, Y. (2017d). Compressive sensing of hyperspectral images via joint tensor tucker decomposition and weighted total variation regularization. *IEEE Geoscience and Remote Sensing Letters*, 14(12):2457–2461.

Wang, Z., Bovik, A. C., and Lu, L. (2002). Why is image quality assessment so difficult? In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–3313. IEEE.

Wu, Y., Rosca, M., and Lillicrap, T. (2019). Deep compressed sensing. *arXiv preprint arXiv:1905.06723*.

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 *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 518–525.

Yang, C., Everitt, J. H., Du, Q., et al. (2010). Applying linear spectral unmixing to airborne hyperspectral imagery for mapping yield variability in grain sorghum and cotton fields. *Journal of Applied Remote Sensing*, 4(041887):041887.

Yu, G. and Sapiro, G. (2011). Dct image denoising: a simple and effective image denoising algorithm. *Image Processing On Line*, 1:292–296.

Yuan, Z., Wang, H., and Wang, Q. (2019). Phase retrieval via sparse wirtinger flow. *Journal of Computational and Applied Mathematics*, 355:162–173.

Yuan, Z., Wang, Q., and Wang, H. (2017). Phase retrieval via sparse wirtinger flow. *arXiv preprint arXiv:1704.03286*.

Zhang, C. and Chen, X. (2009). Smoothing projected gradient method and its application to stochastic linear complementarity problems. *SIAM Journal on Optimization*, 20(2):627–649.

Zhang, H., Chi, Y., and Liang, Y. (2018). Median-truncated nonconvex approach for phase retrieval with outliers. *IEEE Transactions on Information Theory*, 64(11):7287–7310.

Zhang, H. and Liang, Y. (2016). Reshaped wirtinger flow for solving quadratic system of equations. In *Advances in Neural Information Processing Systems*, pages 2622–2630.

Zhang, H., Zhai, H., Zhang, L., and Li, P. (2016). Spectral–spatial sparse subspace clustering for

hyperspectral remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(6):3672–3684.

Zhang, H., Zhou, Y., Liang, Y., and Chi, Y. (2017). A nonconvex approach for phase retrieval: Reshaped wirtinger flow and incremental algorithms. *The Journal of Machine Learning Research*, 18(1):5164–5198.

Zhang, J. and Ghanem, B. (2018). Ista-net: Interpretable optimization-inspired deep network for image compressive sensing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1828–1837.