UNIVERSITY OF PADOVA
DEPARTMENT OF INFORMATION ENGINEERING

DOCTORAL THESIS

Learning For Computational Image Sensing

Author:
Mazen MEL

Supervisor:
Prof. Pietro ZANUTTIGH
Coordinator:
Prof. Fabio VANDIN

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

PhD. Course: Information Engineering
Curriculum: Information Science and Technologies
Cycle: XXXVII

This thesis is written with the financial contribution of the Department of Information Engineering & Sony Europe B.V.

Declaration of Authorship

I, Mazen MEL, declare that this thesis titled, "Learning For Computational Image Sensing" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"An expert is a person who has made all the mistakes that can be made in a very narrow field."

Niels Bohr

Abstract

This thesis deals with two challenging computational imaging problems: spectral and phase imaging. Standard image sensors heavily down-sample spectral information via color filtering and limited spectral sensitivity and are sensitive only to light intensity, which may result in the loss of valuable information related to object composition or the direction of incoming light rays. Consequently, vision and sensing applications such as spectral imaging, interferometry, and 3D imaging are hindered. The aim of this work is to provide ways to computationally recover this information from compressed, multiplexed, and decimated measurements captured using ad-hoc devices. To this end, two learning-based approaches are proposed.

The first approach tackles the problem of hyper-spectral image reconstruction from compressed sensor measurements captured using a CTIS prototype, which is a snapshot imaging device that captures three-dimensional hyper-spectral data cubes as two-dimensional multiplexed signals. Computational post-processing is then needed to recover the latent data cube. However, iterative algorithms typically used to solve this task require large computational resources as the CTIS system matrix is quite wide and can become intractable with a higher spatial resolution of the input measurement. Furthermore, these approaches are very sensitive to the assumed systems and noise models. In addition, the poor spatial resolution of the 0^{th} diffraction order image limits the usability of CTIS in favor of other snapshot spectrometers, even though it enables higher spectral resolution. A novel approach, dubbed Hyper-Spectral and Super-Resolution Network (HSRN) and its subsequent variant HSRN+ are proposed in this regard to recover high-quality hyper-spectral images leveraging complementary spatio-spectral information scattered across the sensor image, furthermore a reconstruction capability beyond the spatial resolution limit of the 0^{th} diffraction order is achieved with quasi real-time performance.

The second approach focuses on Quantitative Phase Imaging (QPI) and recovers a high-quality complex light field from in-line holographic measurements, the phase of which can be used to reveal the contrast in transparent and extremely thin microscopic specimens. Despite the limitation of image sensors, which detect only light intensity, phase information can still be recorded within a two-dimensional interference pattern between two distinct light waves. This work introduces HoloADMM, an interpretable, learning-based approach designed for in-line holographic image reconstruction. HoloADMM enhances the phase imaging capability with spatial image super-resolution, offering a versatile framework that accommodates multiple illumination wavelengths and supports extensive refocusing ranges with up to 10 μm precision. HoloADMM can achieve a substantial improvement in reconstruction quality over existing methods and demonstrates effective adaptation to real holographic data captured by a custom-made DIHM prototype.

Sommario

Questa tesi tratta due importanti problemi di imaging computazionale: imaging multispettrale e phase imaging. I sensori di immagini standard acquisiscono solo una piccola parte dell'informazione spettrali a causa del filtraggio del colore e della limitata sensibilità spettrale, e sono sensibili solo all'intensità luminosa, il che può comportare la perdita di informazioni preziose relative alla composizione della superficie dell'oggetto o alla direzione dei raggi di luce in arrivo. Di conseguenza, applicazioni di visione e rilevamento come l'imaging spettrale, l'interferometria e l'imaging 3D sono difficoltose. L'obiettivo di questo lavoro è fornire modi per recuperare computazionalmente queste informazioni da misurazioni compresse, multiplexate e decimate ottenute tramite dispositivi ad-hoc. A tal fine, vengono proposti due approcci basati su machine learning.

Il primo approccio affronta il problema della ricostruzione di immagini iperspettrali a partire da misurazioni ottenute utilizzando un prototipo CTIS, un dispositivo di imaging istantaneo che cattura dati iperspettrali tridimensionali come segnali multiplexati bidimensionali. È necessario un post-processing computazionale per recuperare l'informazione iperspettrale. Tuttavia, gli algoritmi iterativi tipicamente utilizzati per risolvere questo problema richiedono grandi risorse computazionali poiché la matrice del sistema CTIS è piuttosto ampia e può diventare intrattabile con una risoluzione spaziale della misurazione in ingresso più alta. Inoltre, questi approcci sono molto sensibili ai modelli di sistema e di rumore assunti. Oltretutto, la scarsa risoluzione spaziale dell'immagine dell'ordine di diffrazione 0^{th} limita l'utilizzabilità del CTIS a favore di altri spettrometri istantanei, sebbene consenta una risoluzione spettrale più elevata. Gli approcci proposti (Hyper-Spectral and Super-Resolution Network (HSRN) e la sua variante successiva HSRN+) sono capaci di recuperare immagini iperspettrali di alta qualità sfruttando le informazioni spazio-spettrali complementari sparse nell'immagine del sensore. Così facendo è stata raggiunta una capacità di ricostruzione oltre il limite di risoluzione spaziale dell'ordine di diffrazione 0^{th} operando quasi in tempo reale.

Il secondo approccio si concentra sull Quantitative Phase Imaging (QPI) e recupera campi di luce complessi di alta qualità da misurazioni olografiche in-line, la cui fase può essere utilizzata per rilevare il contrasto in campioni microscopici trasparenti ed estremamente sottili. Nonostante la limitazione dei sensori di immagini, che rilevano solo l'intensità luminosa, le informazioni di fase possono comunque essere registrate all'interno di un pattern di interferenza bidimensionale tra due onde luminose distinte. Questo lavoro introduce HoloADMM, un approccio basato su deep learning interpretabile, progettato per la ricostruzione di immagini olografiche in-line. HoloADMM migliora la capacità di imaging di fase con immagini ad alta risoluzione spaziale, offrendo un framework versatile che può gestire svariate lunghezze d'onda di illuminazione e supporta ampi intervalli di rifocalizzazione con precisione fino a 10 μm. HoloADMM è in grado di raggiungere un

miglioramento sostanziale nella qualità della ricostruzione rispetto ai metodi esistenti ed ha dimostrato un'efficace adattabilità ai dati olografici reali ottenuti tramite un prototipo DIHM appositamente realizzato.

Acknowledgements

First of all I would like to thank Prof. Pietro Zanuttigh for offering me this wonderful learning opportunity and for his unwavering support and guidance during the many years he supervised my work not only as a PhD. student but also as an intern and a master thesis student. Prof. Zanuttigh shaped my way of critical thinking and approaching complex problems in the broad field of computer vision.

I would like to thank Sony Europe for supporting my doctoral studies where I have conducted part of my research at their Stuttgart Technology Center. I was fortunate to work with great talents and experts in the field of imaging, particularly Dr. Alexander Gatto, who supervised my work. I would like to thank him for offering me this interdisciplinary opportunity, for his supervision, and for his support during my time in Stuttgart. I would like to extend my gratitude to Paul Springer, who supervised me during the last stretch of my PhD at Sony and from whom I learned a great deal about the intricate connection between hardware and software parts in complex imaging systems. My deepest gratitude to Markus Kamm for his valuable guidance and intriguing ideas and from whom I learned a lot about imaging and optics, to Ralf Müller for his valuable support in the optics lab, and to Jaqueline Kulmann for her dedication and time spent supporting us in data collection tasks. I would like to thank Dr. Gianluca Agresti for his technical support and helpful ideas and for the fruitful conversations and the fun times we had during my visit to Sony, and all other colleagues with whom I shared wonderful time during my PhD: Roberto Franceschi, Taishi Ono, Dr. Simon Amann, Piergiorgio Sartor, Shan Lin, Dr. Christian Sormann, Dr. Prasan Ashok Shedligeri.

I am grateful to my lab colleagues and the journey we shared together during our research and studies and the fun times we had in Padova and the summer schools: Daniele Mari, Francesco Barbato, Donald Shenaj, Giulia Rizzoli, Elena Camuffo, Federico Lincetto, Matteo Caligiuri, Dr. Umberto Michieli, Dr. Marco Toldo, and Dr. Adriano Simonetto.

I would like to thank my close friends, Yassine Bensaad and Adnen Abdessaied for the wonderfull time we spent in Stuttgart and for their great support, Nicola Stecchetti, Paolo Zinesi, and Andrea Beifiori for the fun times we had in Via Zara.

To my best friend Alaeddine Zaghdoud, thank you for the unequivocal support and guidance during these many years, for always believing in me, and for never losing faith.

To my family, my deepest gratitude for always being by my side and for your support.

To all the people who were part of this journey, my deepest gratitude.

Contents

List of Figures

List of Tables

List of Abbreviations

PSF	**P**oint **S**pread **F**unction
HSI	**H**yper **S**pectral **I**maging
CASSI	**C**oded **A**perture **S**napshot **S**pectrometer
CTIS	**C**omputed **T**omography **I**maging **S**pectrometer
MACTIS	**M**ulti **A**perture **C**omputed **T**omography **I**maging **S**pectrometer
HSI	**H**yper-**S**pectral **I**maging
HSRN	**H**yper-**S**pectral and **S**uper-**R**esoluion **N**etwork
HSRN+	**H**yper-**S**pectral and **S**uper-**R**esoluion **N**etwork **+**
QPI	**Q**uantitative **P**hase **I**maging
DIHM	**D**igital **I**n-line **H**olographic **M**icroscope
ADMM	**A**lternating **D**irection **M**ethod of **M**ultipliers

To my parents and grandparents

Chapter 1

Introduction

In this introductory chapter, basic concepts related to light, propagation, and imaging conditions will be introduced. Afterward, the digital imaging pipeline will be discussed, which encompasses different camera modules and processes involved in generating a two-dimensional image. Vision-based sensing applications and the inverse ill-posed nature of recovering meaningful sensory data from 2D measurements will be discussed with the two primary use cases investigated in this work. Finally, the thesis contributions will be presented in the last section.

1.1 Background and Motivation

In recent years, the field of imaging has undergone significant advancements, particularly in the development of spectral imaging and holographic phase imaging techniques. These advancements have broadened the scope of vision-based sensing applications, which now play a critical role in various scientific, medical, and industrial domains. The primary motivation behind this thesis is to address the challenges associated with these advanced imaging techniques and to contribute to their further development.

Spectral imaging allows for the capture of information across different wavelengths, providing detailed insights into the material properties and chemical composition of objects. Holographic phase imaging, on the other hand, enables the measurement of optical phase shifts, which are crucial for understanding the 3D structure and refractive index variations in transparent or semi-transparent samples.

1.2 Light Representation

The representation of light varies depending on the type of optical simulation and the specific phenomena being investigated. Light can be described using three main frameworks: ray optics, wave optics, and vectorial light fields.

Ray optics, or geometric optics, is the simplest model, treating light as rays that travel in straight lines and change direction according to the laws of reflection and refraction. This approach is particularly useful for studying large-scale optical systems, such as lenses and mirrors, where wave effects like diffraction and interference

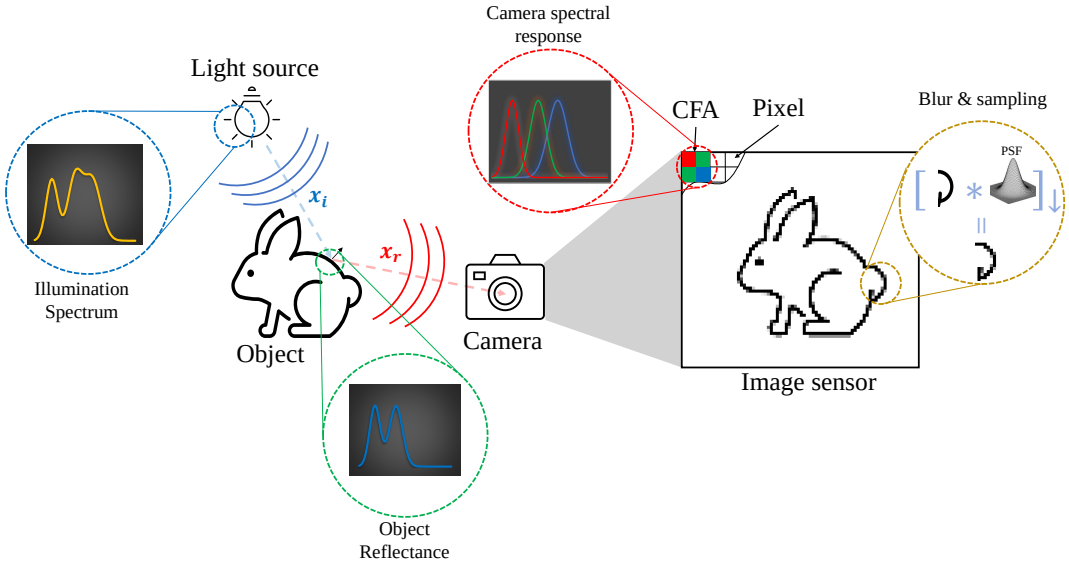


FIGURE 1.1: Detailed schematic of image formation: interactions between object reflectance, illumination, camera spectral response, and image sensor processing.

are negligible. Ray optics will be used later on to describe the imaging condition and optical aberrations in most imaging systems.

However, when dealing with phenomena such as interference, diffraction, or polarization, wave optics becomes necessary. In this framework, light is described as an electromagnetic wave characterized by its wavelength, amplitude, and phase. Wave optics is essential for understanding how light behaves when it encounters obstacles or apertures on the scale of its wavelength.

In the vectorial interpretation, a light field is characterized by electric and magnetic field vectors that oscillate perpendicularly to each other and to the direction of propagation. This vector nature is crucial in understanding polarization effects, birefringence, and the behavior of light in anisotropic media.

In the second part of this thesis, the scalar diffraction theory or the wave nature of light will be considered, a simplified model that approximates the vector nature of light by treating it as a scalar field. This approach is often sufficient for analyzing situations where the polarization effects of light are either negligible or uniform across the field. Here, the light field, denoted by $\mathbf{x}$, is described by its amplitude A and phase Φ:

$$\mathbf{x} = Ae^{j\Phi} \tag{1.1}$$

The intensity I of the light, which is what image sensors measure, is proportional to the squared amplitude of the field, $I = |\mathbf{x}|^2 = A^2$. The phase Φ encodes information about the wavefronts of the light, determining the direction and nature of its propagation.

Standard image sensors, such as CCD or CMOS sensors, are only sensitive to the intensity of light because they cannot directly detect the rapidly oscillating electric

field of light, which typically oscillates at frequencies around 10^{15} Hz. As a result, these sensors cannot capture the phase information of the light field, which is crucial for applications like holography and phase-contrast imaging.

To access phase information, additional techniques such as interferometry are required, where the phase differences between light waves are converted into intensity variations that sensors can detect.

1.3 Image Formation

To fully grasp how digital images are formed, it's important to understand the image formation model in standard digital cameras and how light interacts with camera optics. This includes the process of capturing and digitizing incoming light via an image sensor. Figure 1.1 depicts a conceptual model of the imaging process, illustrating how various components contribute to the final image captured by a camera. The key elements in this model include the object being imaged, the light source that illuminates the object, and the camera's image sensor. Light sources can be classified based on their spectral properties. Broadband sources, like sunlight or halogen lamps, emit light across a wide range of wavelengths. In contrast, monochromatic sources, such as lasers, emit light at a single wavelength, while quasi-monochromatic sources, like LEDs, emit light within a narrow wavelength band. The light emitted from the source interacts with the object, resulting in a reflected light that carries information about the object's reflectance and spectral properties. This reflected light is then captured by the camera through a lens, which can introduce a point spread function (PSF) that contributes to the blurring of the image. Additionally, the camera's spectral response and the color filter array (CFA) affect the final image by modifying the spectrum and resolution of the captured light. The figure aims to visualize the interplay between these factors—object reflectance, illumination spectrum, camera spectral response, and the blur and sampling effects—and to highlight how they collectively influence the image formation process and produce a degraded image of the scene. The aim of this work is to recover lost information such as light spectrum and phase and restore the perceptual quality of the image by performing denoising and spatial super-resolution thus counteracting the effects of the PSF and the discrete and finite pixel grid sampling.

The following sub-sections will go more in details and guide the reader through the camera optics using: (i) a simplified model of ray optics, which is useful for understanding imaging conditions, spot size, and optical aberrations, (ii) scalar wave theory of light and Fourier optics to explore the diffraction-limited PSF, which defines the physical lateral resolution limit of any imaging device. Then, the sensor chip is presented along with its role in capturing incoming photons, digitization and subsequent noise introduction. Finally, the image signal processing unit is described with its different components that contribute to the final image data to be used or displayed later on.

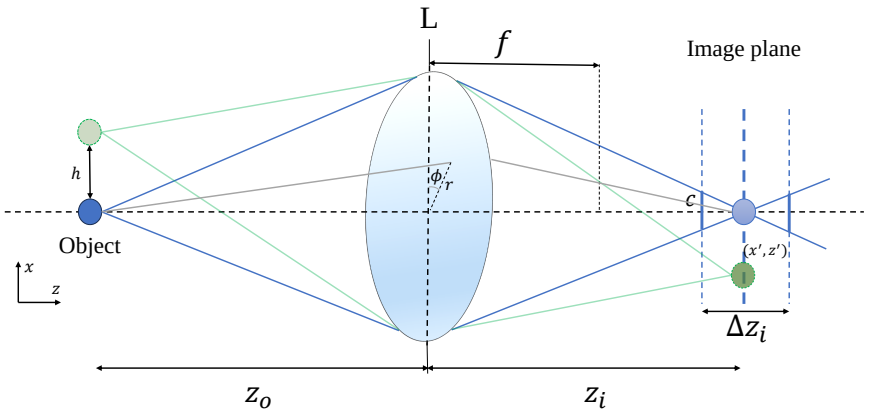


FIGURE 1.2: A thin lens model where light rays are focused by the lens. An object is in focus if the thin lens equation is satisfied. If not, its image would be blurred and spread out across the circle of confusion C.

1.3.1 Thin Lens Model and Optical Aberrations

This simplistic model is widely used in ray optics to derive light transport equations under paraxial regime. A basic thin lens model is shown in figure 1.2. Light rays reflected by an object in the scene reach the lens L with a focal length f_l. The primary function of the lens is to focus light rays coming from an object at a distance z_o in front of the lens into an image at a distance z_i behind it. The thin lens equation is satisfied when all rays coming from point-like sources are focused into a point in the image plane otherwise known also as the imaging condition:

$$\frac{1}{z_o} + \frac{1}{z_i} = \frac{1}{f_l} \tag{1.2}$$

Moving the object away from the focus plane or equivalently moving the image plane, such that the thin lens equation is no longer satisfied, produces a blurry image of said object, known as the circle of confusion. A point is considered in focus if its circle of confusion (C) is smaller than the pixel size of the image sensor. The distance margin within which this condition holds is defined as the Depth Of Field (DOF).

In reality, the thin lens model is almost always violated which means that light rays coming from a single source cannot all converge to the same exact point in the image plane. For instance, the paraxial approximation presumed before does not always hold, as light rays might have large enough angle with the optical axis, those rays do not satisfy the thin lens equation. In real manufactured lenses, surfaces do not perfectly match the designed profile. Furthermore, lenses are dispersive elements as they can only be designed for a single wavelength, all other light rays with different colors will be dispersed differently resulting in chromatic aberrations. Nevertheless, a real camera has a much more complex and sophisticated lens system that minimizes multiple types of optical aberrations. These aberrations are inherent properties of the camera lens and affect how light is focused on the image plane.

They can be mathematically modeled using Seidel's third order correction to the paraxial approximation for an off-axis point situated at $x = h$ and $z = 0$ (refer to figure 1.2):

$$\Delta x' = B_1 r^3 cos(\phi) + B_2 r^2 h(2 + cos(2\phi)) + (3B_3 + B_4) r h^3 cos(\phi) + B_5 h^3 \tag{1.3}$$

$$\Delta z' = B_1 r^3 sin(\phi) + B_2 r^2 h sin(2\phi) + (B_3 + B_4) r h^2 sin(\phi) \tag{1.4}$$

Where $\Delta x'$ and $\Delta z'$ are the deviation from the unique convergence points in the image plane x' and z', (r,ϕ) are point coordinates at the lens plane, B_i , $i = \{1,2,3,4,5\}$ is the strength coefficient of different types of aberrations. Plots shown in figure 1.3 are obtained using code from [SK24].

- **Spherical aberrations (B_1)**: Shown in the first row of figure 1.3 caused by the different focus planes of rays coming from the edge of lens compared to those coming from the inner regions of the lens. Notice that this kind of aberration depends on r^3 and ϕ, the spot diagram on the image plane is a circle with radius propositional to r^3.

- **Coma (B_2)**: Shown in the second row of figure 1.3 caused by imperfections in the lens, and it affects primarily off-axis object points ($h \neq 0$) where the imaged point sources have a tail-like artifacts, the severity of this aberration increases as h increases.

- **Astigmatism (B_3)**: Shown in the third row of figure 1.3 caused by the misfocus between perpendicular rays, e.g., rays with $\phi = 0$ or $\phi = pi$ and those with $\phi = \pm\frac{\pi}{2}$. Leading to an image that appears stretched or distorted along certain axes in addition to blur.

- **Field curvature (B_4)**: Shown in the fourth row of figure 1.3 caused by different focus planes of point across the field of view where if all points were to be in focus, the image plane should be curved.

- **Distortion (B_5)**: Shown in the last row of figure 1.3 it is observable when linear edges exhibit some degree of curvature in the image. The amount of distortion increases non-linearly as a function of the object's distance from the principal point. Notice that it only depends on h^3 so all points will be in focus but their position changes by a factor of h^3.

1.3.2 PSF and Resolution Limit

The simplistic assumptions described in the previous section cannot explain the physical limitation of a real imaging system where point sources of light cannot be focused into points in the image plane due to diffraction limit even when using perfect optics. Therefore, it is necessary to accurately model light behavior and the imaging system using Fourier optics and scalar/wave nature of light instead of

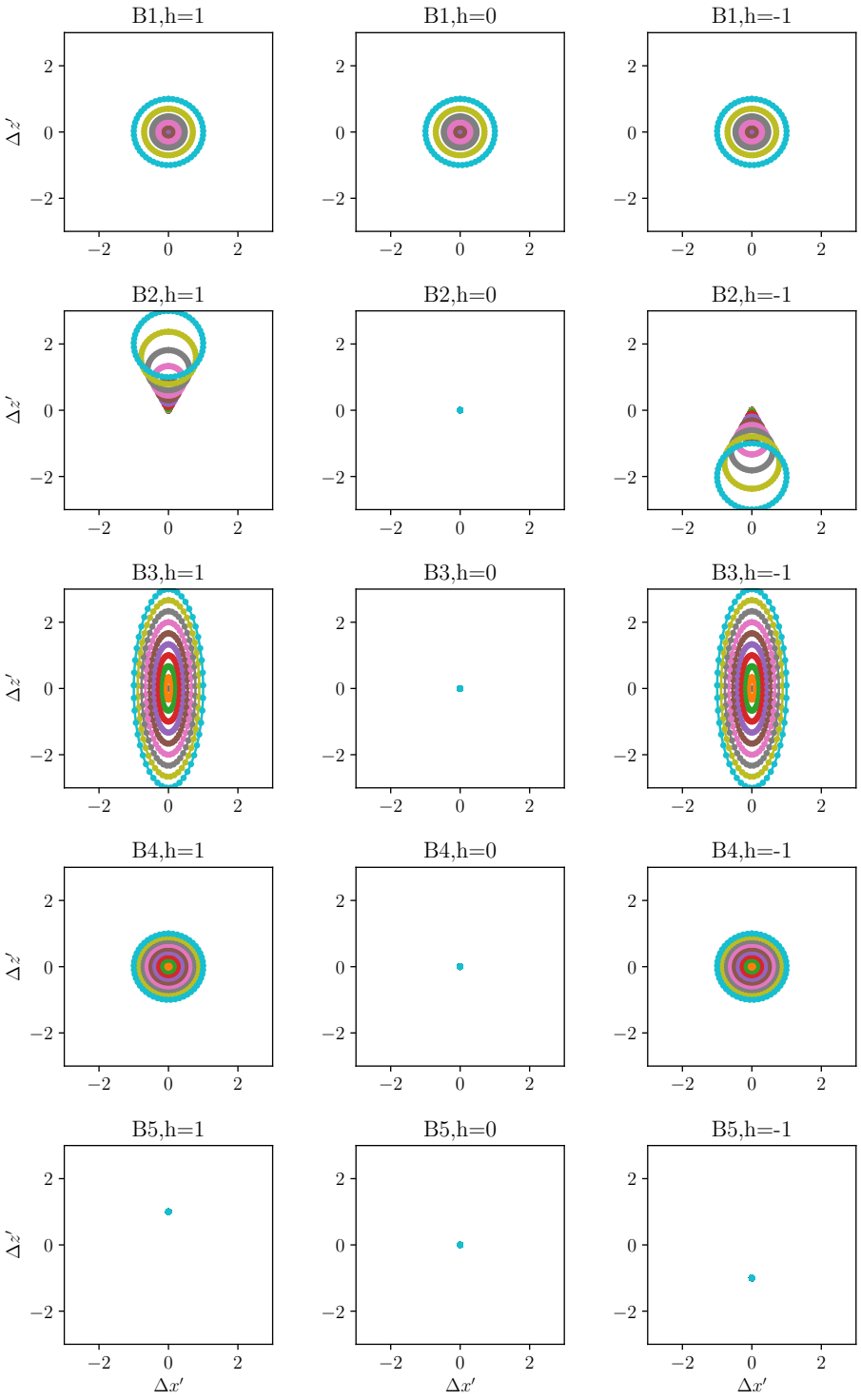


FIGURE 1.3: Optical aberrations caused by the lens, resulting in different types of image misfocus [SK24]. From the first to last rows: (1) spherical aberration, (2) coma, (3) astigmatism, (4) field curvature, and (5) distortion. h is the off-axis vertical displacement.

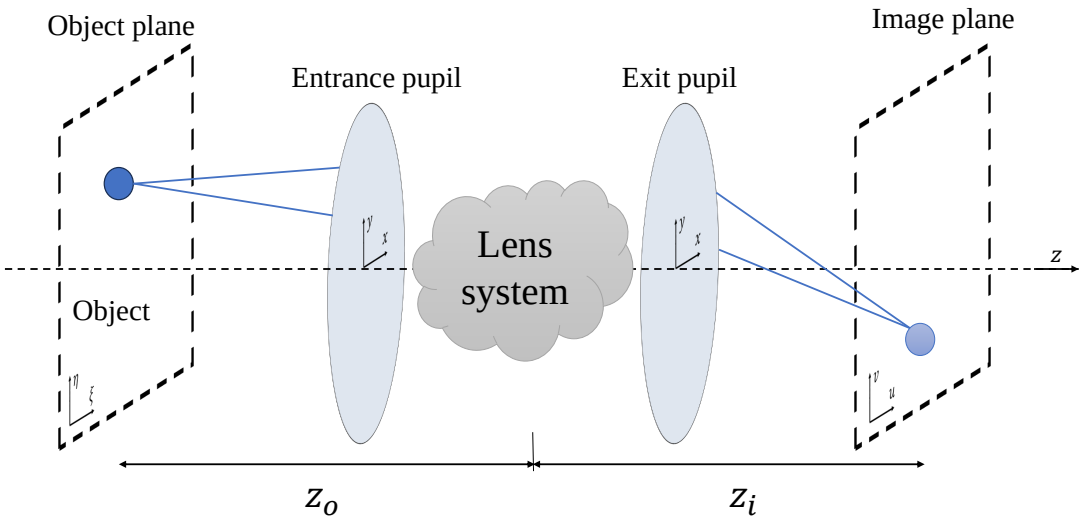


FIGURE 1.4: Generalized imaging system characterized by its entrance and exit pupils.

the geometric ray presentation. To this end, the lens system is viewed as a black box where one would be interested in modeling light waves just at the entrance and the exist planes of such system, known otherwise as the entrance and exit pupils of the system. The reader is encouraged to refer to [Goo05] for more details. The new system schematics is shown in figure 1.4.

Such model takes into account a diffraction-limited imaging system. For instance, light from an ideal point source object passing through a circular aperture would be diffracted at the edges of the aperture, resulting in a diffraction pattern spot on the image plane formed through constructive and destructive interference, known as an *Airy disk*. This is the Point Spread Function (PSF) of the imaging system, defined as the system's response to an ideal point-like source of light. The PSF size limits the size of resolvable object features in the image plane.

Here, a generalized lens system is modeled by entrance and exit pupil planes, where the propagating field is defined when entering and exiting the lens system. As a first simple hypothesis, the object is considered to be illuminated by a coherent and monochromatic light source (e.g., a spatially coherent laser beam). Under this assumption, the image formation model is linear in complex amplitude.

Let $h_c(u, v)$ be the amplitude transfer function of the lens system. The linear forward model can be written as:

$$U_i(u, v) = \iint h_c(u, v; \xi, \eta) U_o(\xi, \eta) d\xi d\eta \tag{1.5}$$

Where U_i is the image amplitude, and U_o is the amplitude distribution emitted by the object. $h_c(u, v)$ can also be seen as the amplitude distribution of a point source of light at the image coordinates (u, v), which is the coherent PSF of the imaging

system. Let $A(x, y)$ be a circular aperture with radius r defined as:

$$A(x,y) = \begin{cases} 1 & \text{if } \sqrt{x^2+y^2} \leq r \\ 0 & \text{otherwise} \end{cases} \tag{1.6}$$

For a point source in focus, $h_c(u, v)$ is expressed as:

$$h_c(u,v) = \frac{C}{\lambda z_i} \iint A(x,y) e^{-j\frac{2\pi}{\lambda z_i}[(u-M\xi)x+(v-M\eta)y]} dxdy \tag{1.7}$$

Where C is a constant factor, λ is the wavelength of the incident light, and $M = -(z_i/z_o)$ is the system's magnification.

In a second, more generalized hypothesis, the illumination is assumed to be incoherent and poly-chromatic light source, which is often the case when capturing images in uncontrolled environments, such as in natural scenes. Thus, it is necessary to generalize the previous expressions to account for an incoherent and poly-chromatic illumination source. Instead of the forward model being linear in complex amplitude, it is now considered linear in intensity distribution. Therefore, the image intensity distribution $I_i(u, v)$ can be expressed as:

$$I_i(u,v) = \kappa \iint |h_c(u-M\xi, v-M\eta)|^2 I_o(\xi/M, \eta/M), d\xi, d\eta \tag{1.8}$$

Where κ is a real constant, I_i and I_o are the intensity distributions of the captured image and the ideal object image, respectively.

As seen from Eq. 1.8, the resulting image is the product of the convolution between the ideal object image and the imaging system's incoherent PSF, which is by definition the square modulus of its coherent counterpart at the in-focus plane:

$$PSF(u,v) = |h_c(u-M\xi, v-M\eta)|^2 \tag{1.9}$$

Now that the incoherent PSF of the imaging system is defined, it is possible to define its Optical Transfer Function (OTF) as follows:

$$\begin{aligned} OTF(f_x, f_y) &= \frac{\iint PSF(u,v) e^{-j2\pi(f_x u + f_y v)}, du, dv}{\iint PSF(u,v), du, dv} \\ &= \mathcal{F}\{\widetilde{PSF}(u,v)\} \end{aligned} \tag{1.10}$$

Eq. 1.10 shows that the OTF of an imaging system with incoherent illumination is the Fourier transform of its normalized point spread function $\widetilde{PSF}$ over the spatial frequencies f_x and f_y.

The Rayleigh criterion is often used to define the spatial resolution limit in diffraction-limited systems. It states that two point sources are considered resolvable when the central maximum of one PSF coincides with the first minimum of the adjacent PSF. The distance at which this occurs depends on the wavelength of light

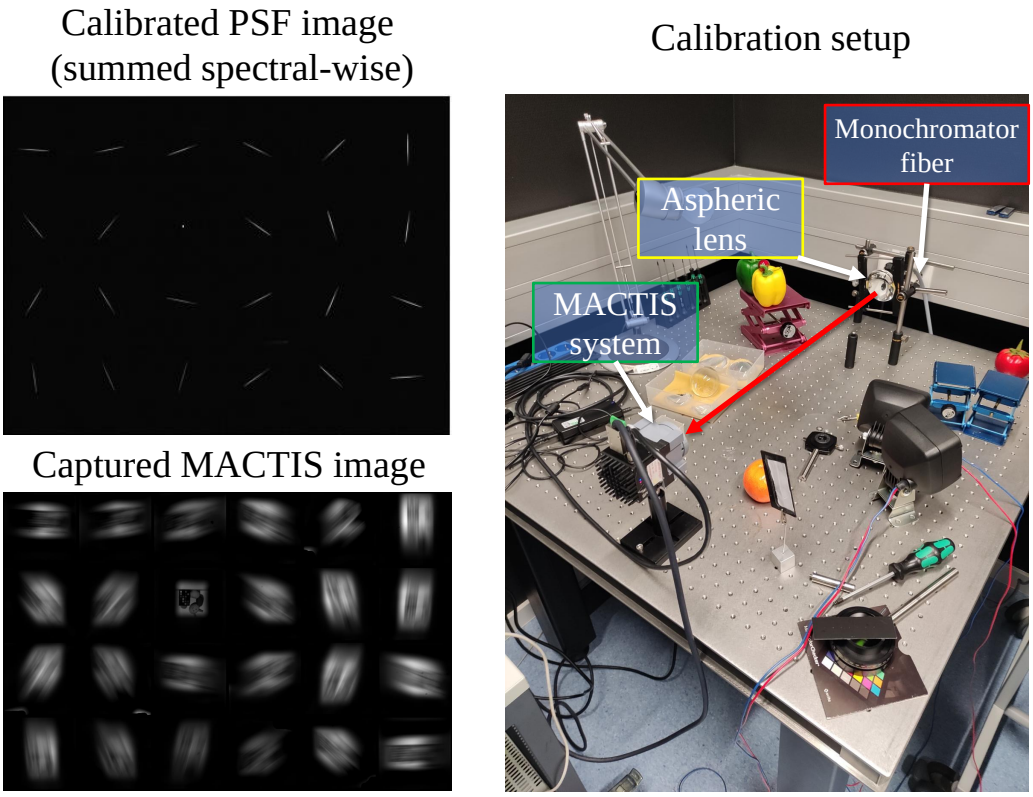


FIGURE 1.5: PSF Calibration setup of a spectrometer system (MACTIS) using an optical fiber connected to a monochromator (right), calibrated PSF image of the system integrated spectral-wise (left).

and the numerical aperture (NA) of the imaging system, and it directly relates to the spot size of the diffracted PSF.

$$\alpha = 1.22 \times \lambda \times \frac{f}{D} \tag{1.11}$$

Where α is the radius of the first dark ring of the *Airy disk* pattern, λ is the wavelength of light, f and D are respectively the focal length and aperture diameter. Two point sources are spatially resolved only when the distance between the centers of their *Airy disk* is equal to α. A lens larger aperture suffer less from diffraction artifacts leading to a small PSF size (small α) which interns leads to higher spatial resolution.

1.3.3 PSF Calibration

Calibrating the PSF of an imaging system is crucial for characterizing and optimizing the system's performance. The PSF describes how a point source of light is spread by the system, capturing the effects of optical aberrations, diffraction, and other factors that influence image quality. The calibration process begins with selecting an appropriate point source of light. This could be a pinhole illuminated by a laser or an LED with a very small aperture. In this work, optical fibers were used in combination

with monochromatic to mimic point sources of light. The point source is then positioned at a known position relative to the system. Once properly aligned, multiple images of the point source are captured under various conditions, such as different wavelengths. The calibration setup used in this work to calibrate a spectrometer device (further details about the system will be discussed in Chapter 2) is depicted in figure 1.5 on the right along with calibration data shown on the left where the wavelength-dependent PSF of the system is shown as smeared spots in the sensor image. Such calibration method enable accurate PSF recovery but at the same time it requires precise alignment of the different optical components of the calibration setup.

In the absence of dedicated optics for system calibration, one can still infer the PSF across the field of view of the system using blind deconvolution techniques using dedicated PSF calibration targets [JSK08; BSA10; Mos+15], the recovery process in this case is purely computational, where the blur kernel is inferred from the camera measurement using some prior knowledge about the shape of such kernel.

1.3.4 Digital Camera

A standard digital image formation pipeline is shown in figure 1.6. After passing through the lens system and aperture, photons are captured by the sensor pixel grid formed by small light sensitive areas known as photodiodes. Photons hit the photodiode's surface exciting electrons within the material, generating a small electric current proportional to the number of incident photons. The stronger the incoming light, the more electron-hole pairs are generated, and thus, the larger the current. This current is then captured and measured as a voltage drop across the diode, which corresponds to the brightness level at that particular pixel. Note that pixels in a CMOS sensor convert generated charges into voltage locally at the pixel level and then forward it to the amplification stage. CMOS sensors are now widely used in modern consumer cameras. The above mentioned process converts the incoming light into an electrical signal, which can then be further processed to form a digital image.

The sensor exposure time controls how many photons are allowed to hit the sensor at capture time and it is controlled by the shutter, situated usually just in front of the sensor. Shutter speed can be set according to the scene dynamics—capturing fast-moving objects requires a higher shutter speed; otherwise, the image would contain motion blur. Note, however, that shutter speed, aperture diameter, and sensor sensitivity (ISO) are generally combined to achieve the desired output.

A schematic of a single pixel in a CMOS sensor is presented in figure 1.7. Since photodetectors are only sensitive to the light intensity hitting them (i.e., the number of received photons), they cannot differentiate between different wavelengths and hence cannot produce color information about the imaged scene. To overcome this issue, an array of color filters is placed on top of the bare sensor, and each photodiode receives only photons of a single wavelength (i.e., those that can pass through

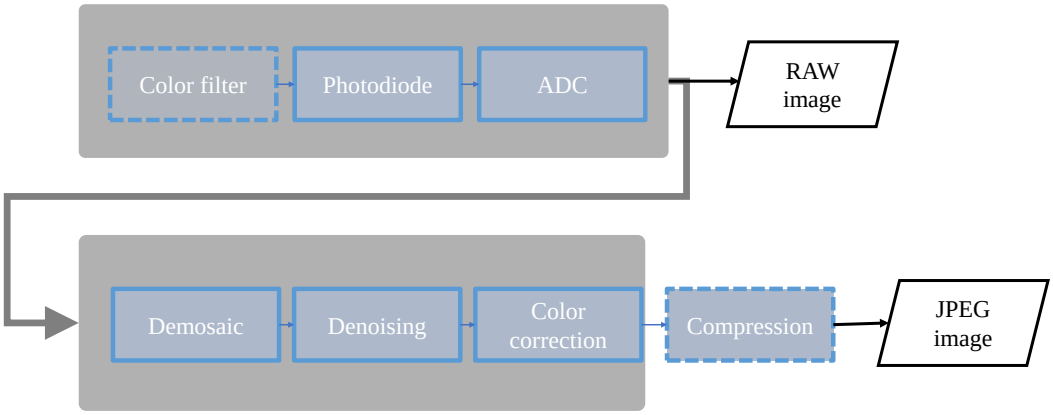


FIGURE 1.6: Modern digital camera's image formation process.

the color filter). A well-known pattern of such Color Filter Arrays (CFA) is the Bayer pattern [Bay76], with a Red-Green-Green-Blue color combination. Half of the grid pixels are green, and the other half is divided equally between red and blue pixels. This is because human eyes are more sensitive to spatial differences in luminance rather than chrominance, with the former being mostly determined by the green color. The process of reconstructing the full RGB channels from the output of the CFA is known as image demosaicing. Worth noting that for many other vision applications, a monochrome sensor is used with no color filters, such sensors are useful for spectral sensing applications which is the main focus of the first part of the thesis.

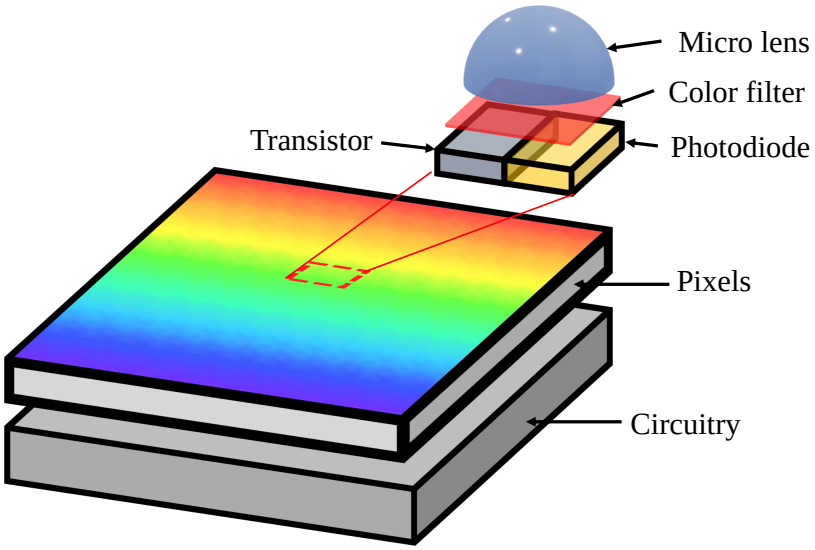


FIGURE 1.7: A single pixel in a CMOS sensor. The photodiode along with the pixel transistor lay below a color filter and a micro lens.

The sensor chip also contains other components, such as micro lenses, which collect light and focus it on the active sensor area of the sensor pixel. Additionally, a

signal amplification and read-out circuit is embedded within the sensor to produce a raw signal.

Before reaching the Analog-to-Digital Converter (ADC), the raw analog signal is amplified by sense amplifiers whose gain can be manually set by choosing the appropriate ISO setting. Otherwise, it can be adjusted automatically by the camera. Note that higher gain leads to the amplification not just of the image signal but also of the noise levels, which can be noticed when capturing photos in low light conditions, resulting in noisy images. The final component in the sensor chip is the ADC, which quantizes the raw analog input signal and outputs a digital signal with a bit resolution of 8, 10, or 16 bits for the RAW output.

The Image Signal Processor (ISP) unit is responsible for raw image post-processing, including demosaicing, white balancing, enhancement of the signal's dynamic range, and image compression to produce the final image in lossy JPEG or lossless PNG formats. However, in this work, mainly raw image data will be used to avoid information loss and distortion that can be caused by those post-processing steps.

1.3.5 Camera Noise Model

Noise in digital cameras refers to the random variations in brightness or color information that can degrade the quality of an image, often appearing as graininess or color speckles. Noise arises from several sources, primarily due to the inherent physical and electronic processes in the sensor and the associated circuitry. The two most common types of noise are **shot noise** and **thermal (or dark) noise**.

Shot noise is due to the quantum nature of light; since photons arrive at the sensor in a random manner, the number of photons captured by each pixel fluctuates, introducing variability in the measured signal. Shot noise follows a Poisson distribution, where the noise σ_{shot} is proportional to the square root of the signal S, represented mathematically as $\sigma_{\text{shot}} = \sqrt{S}$.

Thermal noise, known also as dark shot noise, is generated by the random motion of electrons within the sensor and its circuitry in the absence of light. This noise is proportional to the square root of the sensor temperature and is usually described by the equation $\sigma_{\text{thermal}} = \sqrt{kT}$, where k is the Boltzmann constant and T is the absolute temperature of the sensor.

Another prominent type of noise is readout noise, which occurs during the process of converting the analog signal from the photodiode to a digital signal by the ADC. This noise is often modeled as Gaussian and is independent of the signal.

The overall noise in an image is a combination of these factors and is often expressed as the total noise σ_{total}, which can be approximated as $\sigma_{\text{total}} = \sqrt{\sigma_{\text{shot}}^2 + \sigma_{\text{thermal}}^2 + \sigma_{\text{readout}}^2}$. This total noise impacts the signal-to-noise ratio (SNR), which is a critical factor in determining image quality. High SNR indicates that the signal (actual image data) is much stronger than the noise, resulting in clearer and more detailed images, while low SNR leads to noisy, less

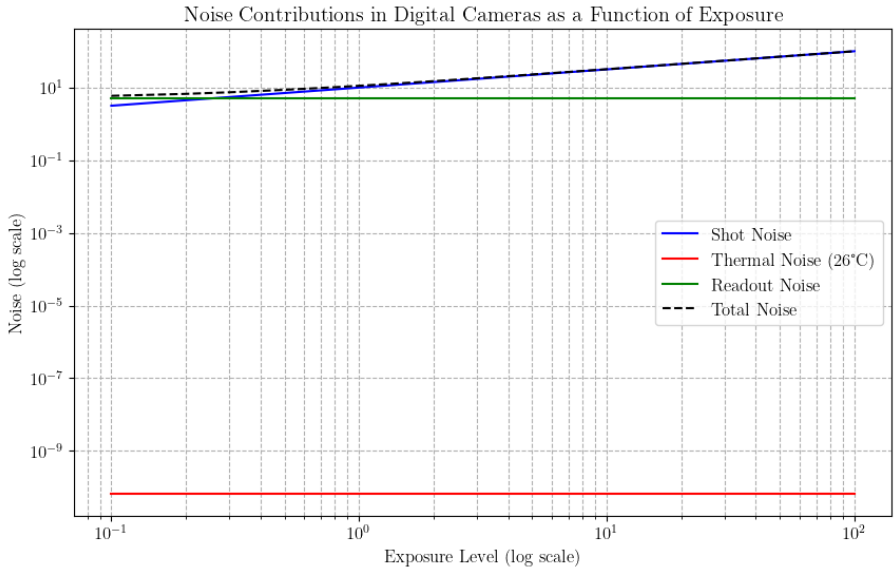


FIGURE 1.8: Contribution of different noise sources.

clear images. The plot shown in figure 1.8 demonstrates the contributions of the different noise sources discussed above with an arbitrary signal and noise variances, read and dark shot noises remain constant as the exposure increases because they are signal-independent, shot noise instead increases with the exposure time and becomes predominant in high-exposure images. Notice that since the contribution of dark shot noise is small it can be neglected in the simulated noise model, in real imaging setups, such noise can be effectively suppressed by subtracting a dark image taken under the same conditions from the actual measurement. In fact, in later sections, the simulated noise model used in the various approaches proposed in this thesis is that of [Foi+08] which considers only read and shot noise sources.

1.4 Inverse Problems in Vision

Inverse problem solving in computational imaging involves reconstructing an unknown image $\mathbf{x}$ from indirect measurements $\mathbf{y}$. Mathematically, the problem can be expressed as:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \epsilon \tag{1.12}$$

Where $\mathbf{A}$ represents the forward model, which describes how the image $\mathbf{x}$ is transformed into the measurements $\mathbf{y}$ and it can be linear as well as non-linear operator, and ϵ denotes the noise or error in the measurements. The goal of inverse problem solving is to estimate $\mathbf{x}$ given $\mathbf{y}$ and $\mathbf{A}$.

In the case of image denoising, the forward operator $\mathbf{A}$ is the identity matrix $\mathbf{I}$ where the expression in Eq. 1.12 is reduced to $\mathbf{y} = \mathbf{x} + \epsilon$, and in the case of image deblurring, the forward model matrix is a Toeplitz block diagonal matrix. In more complex cases like image super resolution, the forward model can be decomposed

into a convolution operator $\mathbf{K}$ with a blur kernel and a down-sampling operator $\mathbf{D}$, where $\mathbf{A} = \mathbf{DK}$. Because the problem of recovering $\mathbf{x}$ is often ill-posed due to the small number of known parameters compared to the unknown ones, directly inverting $\mathbf{A}$, to find $\mathbf{x}$ via inverse filtering is not straightforward due to the presence of measurement noise and when the matrix $\mathbf{A}$ is singular, which is the case in most imaging applications. If the noise is known a priori, then solving a Maximum Likelihood (ML) problem can recover $\mathbf{x}$:

$$\tilde{\mathbf{x}}_{ML} = \arg\max_{\mathbf{x}} p(\mathbf{y}|\mathbf{x}) = \arg\min_{\mathbf{x}} -log(p(\mathbf{y}|\mathbf{x})) \tag{1.13}$$

Where $p(\mathbf{y}|\mathbf{x})$ is the likelihood of observing $\mathbf{y}$ given a true $\mathbf{x}$. The ML solver finds the parameter values that make the observed data most likely. However, it does not use prior information about the parameters and it only relies on the observed data to estimate the parameters. This makes ML purely data-driven, making it sensitive to noise as it would try to fit the noise pattern.

On the other hand, a Maximum A Posteriori Estimation (MAP) is used to tackle the shortcomings of ML by incorporating prior knowledge on the distribution of $\mathbf{x}$. This can be particularly useful when the data is scarce or noisy, as the prior can help guide the estimation process. A MAP estimation problem can be formulated as:

$$\tilde{\mathbf{x}}_{MAP} = \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}) = \arg\max_{\mathbf{x}} p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) = \arg\min_{\mathbf{x}} -log(p(\mathbf{y}|\mathbf{x})) - log(p(\mathbf{x})) \tag{1.14}$$

Where in this case $p(\mathbf{x})$ is the prior distribution on $\mathbf{x}$. In the case of additive Gaussian noise, the MAP framework can be reformulated as a regularized least squares minimization problem:

$$\tilde{\mathbf{x}}_{MAP} = \arg\min_{\mathbf{x}} \frac{1}{2}||\mathbf{y} - \mathbf{A}\mathbf{x}||_2^2 + \Psi(\mathbf{x}) \tag{1.15}$$

Where $\Psi(\mathbf{x})$ is the negative log prior of $\mathbf{x}$ which acts as a regularization term. For example, if the prior is a Gaussian distribution centered around zero, it penalizes large parameter values, preventing overfitting. This regularization effect is similar to adding a penalty term in regularized regression models (e.g., L2 regularization).

In this work, in order to solve the above mentioned MAP estimation problem, learning-based approaches are used. More particularly, neural networks Θ with weights Ω in higher dimensional space are optimized in a supervised manner so that $\tilde{\mathbf{x}}_{MAP} = \Theta_\Omega(\mathbf{y})$ where the network in this case is already trained. During training, the weights are minimized using a gradient descent based optimizer and are trained using a couple $\{\mathbf{y}, \mathbf{x}\}$ of input measurements and target ground truth data drawn from a data distribution:

$$\tilde{\Omega} = \arg\min_{\Omega} \mathcal{L}(\Theta_\Omega(\mathbf{y}), \mathbf{x}) \tag{1.16}$$

Where $\mathcal{L}(\cdot)$ is an objective (or loss) function to be minimized, e.g., it can be the mean squared difference between $\mathbf{x}$ and $\Theta_\Omega(\mathbf{y})$.

1.5 Thesis Contributions

This thesis makes several novel contributions in the field of computational image sensing. In particular, different learning-based approaches are proposed to tackle the image reconstruction problem in snapshot spectral imaging using a tomographic based imaging device and in phase imaging using lens free microscopic devices:

- Chapter 2 is dedicated to spectral imaging where HSRN [MGZ22] and HSRN+ [MGZ24] are introduced which are the outcome of an improved spectral imaging techniques: Wherein novel methods are devised to enhance the resolution and accuracy of spectral images with application to material characterization [Ama+23a].
- Chapter 3 introduces HoloADMM [Mel+24], an approach to tackle image reconstruction quality in microscopic phase imaging which employs holographic techniques to recover light phase distribution using wave interference and generate image contrast of otherwise thin and transparent microscopic specimens.
- Chapter 4 concludes the work carried out in this thesis and present some future research directions concerning both spectral and phase imaging.

These contributions are expected to advance the state of the art in imaging science, providing new tools and methodologies for researchers and practitioners in the field.

1.6 List of Publications

Journals

- Mel, M., Gatto, A., & Zanuttigh, P. (2024). Joint Reconstruction and Spatial Super-resolution of Hyper-Spectral CTIS Images via Multi-Scale Refinement. IEEE Transactions on Computational Imaging.
- Mel, M., Siddiqui, M., & Zanuttigh, P. (2023). End-to-end learning for joint depth and image reconstruction from diffracted rotation. The Visual Computer, 1-17.
- Zimmermann, M., Amann, S., Mel, M., Haist, T., & Gatto, A. (2022). Deep learning-based hyperspectral image reconstruction from emulated and real computed tomography imaging spectrometer data. Optical Engineering, 61(5), 053103-053103.

- Mel, M., Michieli, U., & Zanuttigh, P. (2019). Incremental and multi-task learning strategies for coarse-to-fine semantic segmentation. Technologies, 8(1), 1.

Conference Proceedings

- Zhou, H., Mel, M., Springer, P.,& Gatto, A. (2024). Cross-Net: Joint In-Line Holographic Image Reconstruction and Refocusing. In MICAD 2024
- Mel, M., Springer, P., Zanuttigh, P., Zhou, H., & Gatto, A. (2024). HoloADMM: High-Quality Holographic Complex Field Recovery. In ECCV 2024.
- Amann, S., Mel, M., Zanuttigh, P., Haist, T., Kamm, M., & Gatto, A. (2023, May). Material Characterization using a Compact Computed Tomography Imaging Spectrometer with Super-resolution Capability. In Proceedings of the 6th International Conference on Optical Characterization of Materials, OCM 2023 (pp. 139-148).
- Mel, M., Gatto, A., & Zanuttigh, P. (2022, November). Joint Reconstruction and Super Resolution of Hyper-Spectral CTIS Images. In BMVC (p. 1063).

Patents & Invention Reports

- M. Mel, P. Springer, & P. Zanuttigh. (2024). A Method for Complex Field Recovery from In-line Holographic Measurements with Joint Spatial Super-resolution (filed).
- M. Mel, & A. Gatto. (2023). Enhanced spectral image reconstruction from CTIS image (filed).
- S. Amann, M. Mel, & A. Gatto.(2023). Apparatus and Methods for Computer Tomography Imaging Spectrometry (WO2024083580A1).
- S. Muhammad, & M. Mel. (2021). Camera, Method and Image Processing Method (WO2023001674A2).

Chapter 2

Snapshot Spectral Imaging

2.1 Introduction

Hyper-Spectral Imaging (HSI) techniques enable the capture of multiple spectral bands, extending beyond the standard RGB channels captured by typical imaging devices, as shown in figure 2.1. The spectral data volume can be conceptualized as a three-dimensional cube, with the first two dimensions representing spatial data and the third dimension representing spectral data, containing numerous spectral bands. Throughout this thesis, this data volume will be referred to as a (hyper)-spectral cube, object cube, or (hyper)-spectral images. Rich spectral information is essential for various vision-based sensing applications, including material characterization in sorting and recycling [BLH23], medical image analysis [Kha+18], remote sensing and monitoring [Tek+13], and object detection and tracking [EA+22].

The increasing demand for hyper-spectral data has led to the development of various spectroscopic systems, which can be categorized into two main types: scanning-based devices and snapshot devices: **1) Scanning-based devices** include Whiskbroom spectrometers that can capture a full spectrum pixel-wise or of a handful of pixels at a time, Pushbroom spectrometers that can capture the spectrum of roughly a single line of pixels in each exposure using a slit aperture. Alternatively a tunable color filter (e.g., VariSpec™) can be used to capture a full two-dimensional image for a single wavelength at a time. The aforementioned techniques have to

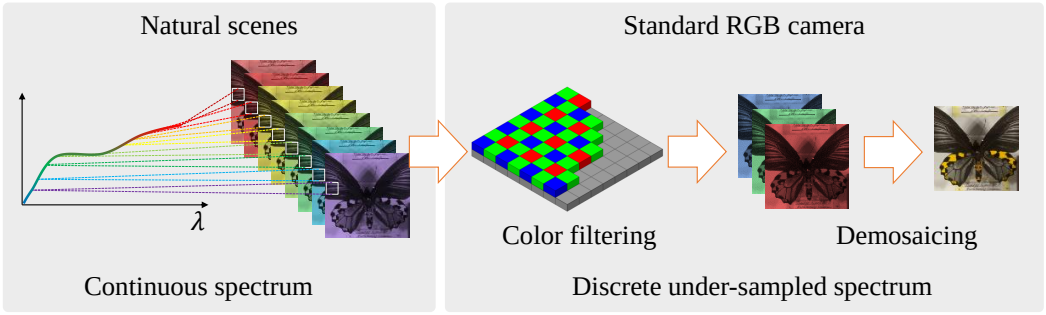
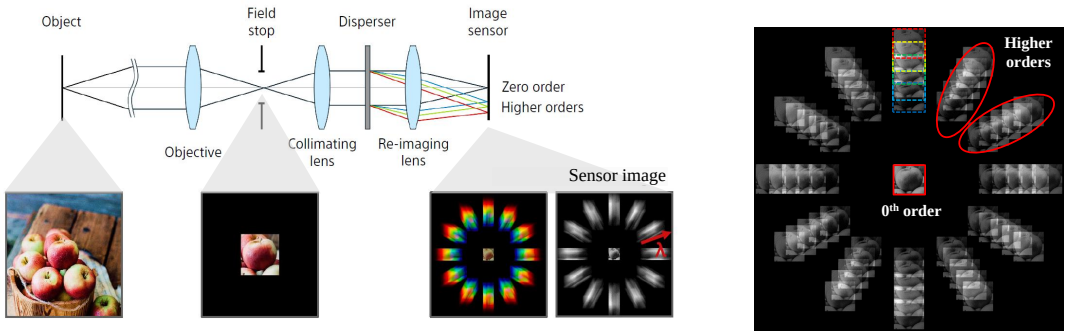


FIGURE 2.1: Continuous light spectrum reflected from the scene is under-sampled through color filtering in RGB sensors. Butterfly image courtesy of [Mon+15].

balance between spectral/spatial resolution and the acquisition time; whenever finer resolution in the spatial or spectral domains is needed, scanning time has to be increased. **2) Snapshot devices** require a single exposure to capture sparse and compressed measurements of the latent object cube wherein spatial and spectral information are multiplexed via some kind of aperture coding and dispersion optical elements as in Coded Aperture Snapshot Spectral Imager (CASSI) systems [Geh+07], Diffractive Optical Elements (DOEs) [Cou60] or eventually using color filter arrays on top of the image sensor [Bil01], similar to a Bayer color filter array but with more color filters than the standard RGB pattern. A comprehensive overview of these techniques can be found in relevant literature which can be found in [HK13].

Compressed measurements from these devices are intermediate representations, requiring post-processing computational algorithms to reconstruct the full three-dimensional object cube. Despite various heuristic and learning-based approaches in the literature, the reconstruction problem remains challenging due to the ill-posed nature of the inverse problem. The linear system is often under-determined, with the number of equations significantly lower than the number of parameters to estimate, necessitating careful prior selection (e.g., sparsity of image data in some transform domain [Dab+08], low rank structure of spectral images across the spectral dimension [Liu+18]) in order to constraint the space of possible solutions. Traditional model-based iterative solvers struggle to recover high-quality spectral images in a reasonable timeframe with relatively small deviation from the latent dense spectral volume even with good image priors. Furthermore, the existing trade-off between spatial/spectral resolution and computational time inherently limits any possible deployment of such solvers and therefore snapshot spectrometers in real-time sensing applications. To this end, learning-based alternatives can address those shortcomings specifically with respect to inference speed and overall reconstruction quality but lack robustness towards new unseen data with large domain gaps. This works deals primarily with Computed Tomography Imaging Spectrometer (CTIS) devices the concept pf which was independently invented by Takayuki Okamoto [OY91] in 1991 and by Theodor V Bulygin [BV92] in 1992.

CTIS is a snapshot spectrometer that rely on dispersive optics like diffraction gratings to separate light based on its wavelength on project it onto a two-dimensional pixel array. The working principle of a CTIS systems is depicted in figure 2.2 where light coming from the scene is first imaged by an objective lens onto the system aperture (field stop) and then a collimating lens is used to parallelise light rays which are then dispersed by a diffraction grating into zero as well as higher order projections and imaged by a monochromatic image sensor. In this way, spectral information of the three-dimensional object cube can be encoded as spatio-spectral multiplexed signals in a two-dimensional imaging medium. The main advantage of CTIS systems is the possibility of achieving very compact form factor allowing such technology to be used in small mobile devices with size and



(A) (B)

FIGURE 2.2: (A) Schematics of a CTIS imager, (B) Sensor measurement of a discretized spectral cube.

weight restrictions.

Even though CTIS is capable of capturing spectral information within a single exposure, recovering the latent object cube form the intermediate two-dimensional measurement is challenging and requires solving an ill-posed inverse problem which stems from the missing cone issue in Fourier Slice Theorem [HB81] as in most other Computed Tomography (CT) based imaging systems. The theorem states that the two-dimensional Fourier transform of each projection within the sensor image is equal to a plane through the three-dimensional frequency representation of the latent object cube. Due to the limited number of projections and the low diffraction efficiency, the full frequency representation cannot be recovered as that would require an infinite number of projections. Consequently, it is impossible, from a mathematical point of view, to recover that missing information since there is now way one can sample within the missing cone with conventional imaging senors. However, satisfactory results can still be achieved using appropriate spectral and spatial prior information about the scene.

In this work, multiple learning-based approaches are proposed to address the shortcomings of CTIS and to tackle the hyper-spectral image reconstruction problem from CTIS measurements beyond the limited spatial resolution of the 0^{th} diffraction order image. The objective is to achieve a quasi-real-time high spatial resolution reconstruction capability. A first approach introduces *P2Cube* [Zim+22] to tackle the spectral reconstruction problem using end-to-end supervised learning, the second further builds upon *P2Cube* and introduces Hyper-Spectral and Super-Resolution Network (HSRN) [MGZ22], a network capable of recovering hyper-spectral images with high spatial resolution from CTIS sensor measurements and it is partially inspired by the conventional Filtered Back-Projection (FBP) algorithm used in most CT scan-based computational reconstruction methods. Finally, HSRN+ [MGZ24] is proposed as an enhanced version of its predecessor HSRN: it features a multi-scale refinement architecture and it is able to reconstruct the three-dimensional spectral cube in a coarse to fine manner. It is targeted towards higher reconstruction quality

from real measurements captured by three different CTIS prototypes that were custom built. Furthermore, a large scale spectral dataset was collected to enable hyper-spectral image reconstruction and semantic segmentation simultaneously. The main contributions of this first part of the thesis can therefore be summarized as the following:

- A simple yet efficient network architectures capable of reconstructing spatially super-resolved object cubes from CTIS measurements in real-time (up to 30 fps for a cube of size $400 \times 400 \times 31$ voxels in the case of *HSRN*) and quasi-real-time performance for *HSRN+*.
- Joint approaches for hyper-spectral image reconstruction and spatial super-resolution from CTIS sensor measurements exploiting aliased information scattered across the image sensor with a novel multi-scale learning framework for incremental image refinement.
- Extensive experimental studies have been conducted to validate models' reconstruction performance on synthetic as well as real CTIS data using three different prototypes achieving significantly better reconstruction quality than the current state-of-the-art.
- Proof of concepts using miniaturized and custom-made CTIS imagers.
- A new large scale dataset for spectral reconstruction and semantic segmentation with high resolution hyper-spectral images and high quality manually annotated segmentation maps that enable end-to-end learning and model validation for material characterization tasks.

2.2 Prior Art

2.2.1 HSI Devices

Early spectrometers were predominantly scanning devices such as pushbroom [PE87], whiskbroom [Bru+06], and tunable color filter cameras [HSB02] which are capable of capturing images with high spatial and spectral resolution but at the same time they are fairly large and cumbersome devices incorporating multiple moving parts and requiring long acquisition times. Owing to the quick advancements in compressive sensing and deep learning, snapshot spectrometers or even conventional RGB cameras [Sim+21; Jia+17] became widely used to capture dense spectrum of dynamic scenes. CASSI systems [Wag+08; Kit+10] stand out as one of the most used devices for HSI with numerous reconstruction algorithms that have been developed thus far to process CASSI measurements. However, these systems offer poor image quality as the spatial resolution is significantly degraded due to the use of coded aperture masks. In addition, its spectral resolution is limited by the sensor pixel pitch along with the non-linear dispersion introduced by the prism

leading to a trade-off between spatial and spectral resolution. Alternatively, in a CTIS system [OY91; BV92; HKW12] light is dispersed into multiple tomographic projections via a Diffractive Optical Element (DOE) forming multiple projections of the latent object cube on the image sensor, even though the spatial resolution of such projections is low leading to sub-optimal use of the sensor area, CTIS provides greater spectral resolution due to its high angle of parallel projections allowing to resolve higher number of spectral bands. Indeed, CTIS practical applicability is reduced by the poor spatial resolution of its 0^{th} diffraction order which determines the resolution of the reconstructed hyper-spectral image [Dou+20; Dou+21; HWC21]. Furthermore, no previous work has tackled this problem so far, at least from a computational point of view. In this work, sub-pixel displacements present in higher diffraction orders are exploited to perform image super-resolution of hyper-spectral cubes with up to $\times 6$ the resolution of the 0^{th} order image thus paving the way for more research into CTIS technology.

2.2.2 Compressive Spectral Imaging

Recovering dense three-dimensional hyper-spectral cubes from compressed two-dimensional sparse measurements is ill-posed. Iterative solvers have been proposed to address the inverse reconstruction problem from several coded inputs depending on the type of spectrometer used to acquire such measurements. Several approaches in the literature relied on carefully selected image priors in a Maximum A Posteriori (MAP) estimation framework tailored most of the time for CASSI systems. IST [DDDM04] and TwIST [BDF07] incorporated a Total Variation (TV) norm regularization term to encourage sparsity of the latent hyper-spectral image in some transform domain. [Liu+18] introduced DeSCI, which uses a weighted nuclear norm regularizer to solve a rank minimization problem exploiting the low rank structure of the three-dimensional latent object cube along the spectral dimension formed by non-local similar patches. Such MAP problem can be solved also via variable splitting techniques such as the Alternating Direction Method of Multipliers (ADMM) [Boy+11] or the Half Quadratic Splitting (HQS) [GY95] method which are used to decouple the data fidelity and prior terms in the energy function subject to minimization, the prior/regularizer can be solved by a plug-and-play image denoiser module using off-the-shelf powerful denoisers such as BM3D [Dab+07] or more recently pre-trained convolution neural network as in [Zhe+21]. Even though impressive performance has been achieved using model-based approaches, the required computational time complexity is still quite large [Men+21]. In an attempt to combine the interpretability and flexibility of model-based approaches with the reconstruction speed at inference time and the large representation capability of deep convolutional neural networks, optimizer-based unrolled network architectures have been introduced in [Zha+21] which combines the best of the two worlds to tackle the reconstruction problem of object cubes from CASSI measurements.

Hyper-spectral images from CTIS measurements are usually obtained by the Expectation-Maximization (EM) solver that is predominantly used for reconstruction [Vol00] in most computed tomography-based acquisition systems. The EM is a Maximum-Likelihood (ML) approach that cannot handle priors and is very sensitive to the presumed noise and system models which can be easily inaccurate in real imaging scenarios leading to sub-optimal performance and poor reconstruction quality. Modified versions of EM aimed at introducing prior knowledge into the optimization framework incorporating some other constraints such as low rank and superiorization have been proposed in [Li+18] and [HWC21]. A recent GPU accelerated EM variant was introduced by [WBH20] and reached significant speedup in reconstruction time but at the expense of a poor spatial resolution, the authors exploited the spatial shift invariance of the system matrix which assumes that shifting the object cube by a certain number of pixels the diffraction pattern at the sensor plane would be shifted by the same amount. Furthermore, under this assumption [HDS07] proved that one can diagonalize much of the system matrix in Fourier domain achieving a computational speedup by a factor of 4000. One can argue that learning-based approaches, thanks to their vast representation capability and fast inference time once trained, are nowadays a very good option for real-time high-quality image reconstruction from CTIS measurement that could achieve a good trade off between spatial and spectral resolution. An approach of this family is that of [Hua+22], who proposed a straight-forward end-to-end learning approach that learns a mapping beween CTIS measurements and ground truth object cubes in a supervised manner through a multi-branch CNN as an analogy to ensemble learning. [Ahl+22] introduced a hybrid approach exploiting a CNN followed by an EM solver where the network provides an initial estimate for EM, the contribution of such hybrid workflow is subject to how sensitive the EM is to changes in the initial guess as the optimization problem is linear and thus the algorithm would likely converge to the same result regardless of the choice of the initial guess. [Zim+22] implemented an initial reshaping layer enabling 3D processing of high dimensional input data to account for spatio-spectral correlations within multiple higher diffraction orders which is then followed by a U-Net like architecture [RFB15] used to refine the estimated object cube. The work of [WC23] used a cGAN-based architecture conditioned on the 0^{th} diffraction order image. Unlike all previous approaches, for the first time the spatial resolution issue for CTIS has been addressed from a computational point of view in the works discussed in this thesis [MGZ22] and [MGZ24]. The proposed approaches reconstruct a spatially super-resolved object cube with up to $\times 4$ the resolution of the 0^{th} order image on synthetic CTIS data and by factors of $\times 6$ and $\times 2$ on two different real data sets captured by different CTIS prototypes. Superior reconstruction capability has been demonstrated with respect to current state-of-the-art with a lightweight model prompting lower inference time compared to other competitors. Similarly, in [Yua+23] the authors used a high-resolution RGB camera in combination with a CTIS system to produce a spatially super resolved

image using spatio-spectral pixel interpolation between the low-resolution EM output and the high-resolution RGB image. In contrast, the approaches proposed in this thesis need just a single image sensor allowing for compact systems to restore a high-quality object cube.

2.2.3 Image Super-Resolution

Image super-resolution aims at recovering a high spatial-resolution image from decimated and noisy sensor measurements. It require solving an inverse ill-posed problem given single or multiple measurements wherein, unlike simple image interpolation, the aim is to recover high-frequency components from aliased low frequency data. Earlier works such as that of [Yan+10] exploit a sparse representation in combination with dictionary learning to map low resolution sparse signals to their high resolution counterparts. Similarly, [GBI09] used an example-based super-resolution approach exploiting non-local self-similarity in natural images. Recent approaches are instead mostly based on deep learning models due to their high capacity to learn complex non-linear mappings from low-resolution to high-resolution image spaces and effectively reconstruct visually appealing high frequency details [Lim+17; Led+17]. In burst imaging pipelines multi-frame "true" image super-resolution exploits information provided by distinct frames in the form of aliasing which is caused by relative camera and/or object motion during exposure with sub-pixel accuracy. Complementary information present in multiple frames is combined via image registration and mapped into a higher resolution pixel grid with high spatial fidelity to the latent high-resolution image. Such methods can be divided into model-based approaches [Wro+19; Li+10] and deep learning-based ones such in [Dud+22] where significant performance gains have been achieved with high resolution factors. In this work, the proposed spatial super-resolution modules are inspired by multi-frame image super-resolution approaches. However, due to the particularity of CTIS measurements, such approaches must be adapted to take into account the structure of the sensor image.

2.2.4 Multi-Scale Learning

Multi-scale learning in the form of hierarchical representation of multi-resolution feature maps is used as a way of producing high-quality outputs starting from a coarse and initial reconstruction and performing intermediate incremental refinement steps leveraging prior coarse knowledge. [EPF14] used a multi-scale deep network for depth refinement from a single input image and propagated coarse depth prediction via shortcut connections to a subsequent fine-scale network by a simple channel-wise concatenation. [DRS20] used a multi-scale approach in the task of non-blind image deblurring where coarse feature maps are incrementally refined through a multi-stage network architecture starting from low up to high resolution space. In

this work such approach is adapted to spectral image reconstruction and spatial image super-resolution and it shows that multi-scale learning in this context eases the reconstruction burden on the network especially when the spatial resolution scale increases as coarse spectral information is propagated through the network via feature level concatenations prompting the network to learn to reconstruct fine spatial details in the final predicted spectral image.

2.2.5 Spectral Image Segmentation

Material characterization benefits from spectral imaging as the task of classifying objects based on their spectral signatures becomes significantly easier when dense spectral information is available at higher spectral resolution compared to standard RGB data with only three color primaries. Multiple works dealt with semantic image segmentation using spectral data as input: in satellite imaging and remote monitoring [YWL13], in industrial and recycling sorting-based applications [Bäc+23], and food inspection [Dja+23]. The lack of sufficiently large high-quality (spectral and spatial-wise) annotated data still limits the development of new approaches for semantic image segmentation from spectral data. This work further contribute to this field by providing an experimental dataset designed to assess and validate the performance of different methods on the tasks of spectral image reconstruction and semantic segmentation for the benefit of material characterization tasks.

2.3 Methodology

2.3.1 CTIS Image Formation Model

The CTIS image sensor features a dispersion-free sharp central image which is the 0^{th} diffraction order of the DOE surrounded by higher order diffraction images: where the 3D object cube is first dispersed by the DOE, i.e., spectral bands are shifted spatial-wise with respect to each other, and then integrated spectral-wise by the image sensor to end up with a smeared and multiplexed two-dimensional projection. The underlying image formation model for CTIS can be written as:

$$g = \mathbf{H}f + \epsilon \tag{2.1}$$

where $g \in \mathbb{R}^{(MN)\times 1}$ and $f \in \mathbb{R}^{(HW\Lambda)\times 1}$ are, respectively, the vectorized sensor image with spatial dimensions $M \times N$ pixels and the latent object cube to be reconstructed with Λ spectral channels and spatial dimensions of $H \times W$ pixels (see figure 2.9). The 0^{th} diffraction order image dimensions are roughly 10% of those of the image sensor. $\mathbf{H} \in \mathbb{R}^{(MN)\times(HW\Lambda)}$ is the system matrix that maps each pixel from the object space to its counterparts in the image space (basically it contains the PSFs for each wavelength and for each projection), and ϵ is an additive noise term. In order to better reproduce actual sensor noise, a combination of additive Gaussian noise

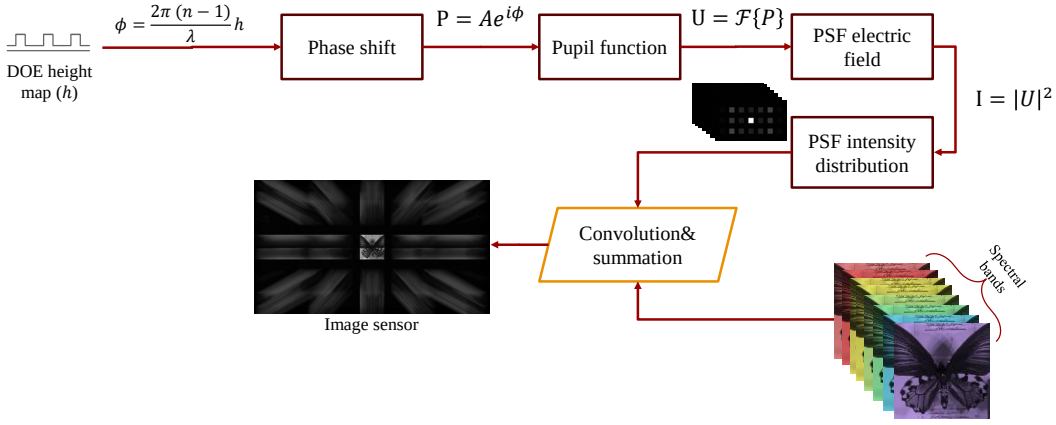


FIGURE 2.3: Fourier optics-based CTIS data simulation pipeline. Butterfly image from [Mon+15]

and signal-dependent shot noise sources can be used in simulating synthetic CTIS measurements.

2.3.2 CTIS Data Simulation

Fourier optics principles are used to simulate CTIS sensor images via wave field propagation (see figure 2.3). Starting from the height map h of the DOE with a refractive index $n = 1.5$ and designed for a reference wavelength $\lambda^* = 550$ nm (i.e., it creates a phase shift of π at λ^*), the phase shift Φ_λ introduced by such DOE on an incident planar wave front is obtained by:

$$\Phi_\lambda = \frac{2\pi(n-1)}{\lambda} h \tag{2.2}$$

The angular spectrum method is then used to propagate the field from the object plane to the sensor plane by applying the Fourier transform of the pupil function P. The PSFs for each wavelength are obtained by calculating the squared modulus of the field at the sensor plane via:

$$PSF_\lambda \propto \left| \mathcal{F}\{A \cdot e^{i\lambda^* \Phi_\lambda}\} \right|^2 \tag{2.3}$$

Where $\mathcal{F}$ is the Fourier Transform operator and A is a binary mask defining the aperture area. The PSF has a large support spanning the total image sensor area and is convolved with ground truth object cube f interpolated across the spectral dimension in order to get as many spectral bands f_λ as possible. All convolution results of the spectral PSF with the corresponding spectral bands are summed together to end up with a two-dimensional CTIS sensor image:

$$g = \sum_\lambda PSF_\lambda * f_\lambda \tag{2.4}$$

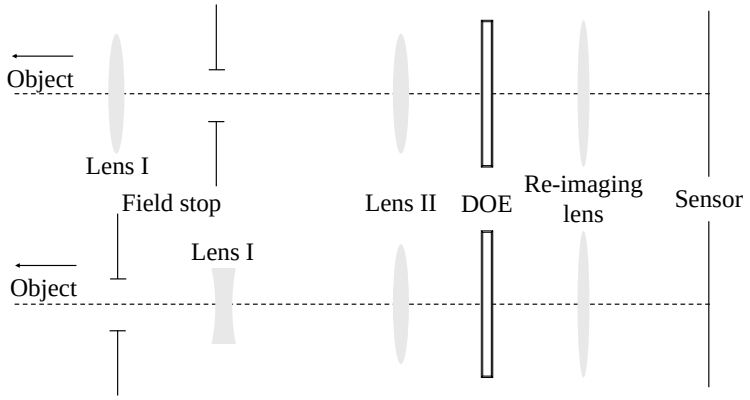


FIGURE 2.4: Top: *Keplerian* CTIS design where the field stop is placed on the back focal plane of the imaging lens (Lens I). Bottom: *Galilean* design where the field stop is placed in front of Lens I and Lens II where the two lenses act as a beam expander.

The obtained sensor image in the simulations has 14 higher diffraction orders surrounding the dispersion-free 0^{th} order in a 3 by 5 formation: each higher diffraction order is basically the ground truth hyper-spectral cube smeared across the spatial dimension following the projection angle.

2.3.3 Real CTIS Data Acquisition

Real CTIS Prototypes: Experiments on real world data were carried out using two different CTIS prototypes, a third prototype dubbed MACTIS was later used to test a new system design and it is discussed in details in Section 2.5.8. Table 2.1 shows the detailed specifications of the two first setups, i.e.: (i) A full-frame CTIS prototype with a *Keplerian* design is used where the field stop is placed in between the imaging lens (lens I) and the collimating lens (lens II) which in turn is followed by the DOE as illustrated in figure 2.4. Such system provides good 0^{th} order image quality with reduced optical aberrations, however the overall form factor of the device is large since the beam expander requires that lens I and II be separated by the sum of their focal lengths, which hinders its usefulness in outdoor or mobile applications. (ii) Alternatively, a *Galilean* beam expander can be used to significantly reduce the form factor of the system [Ama+23a] where the field stop is placed in front of a lens combination featuring two stacked negative lenses and a positive one acting as a beam expander, this allows for a smaller form factor since lenses I and II must be separated by the difference of their focal lengths, the schematic in figure 2.4 illustrate such configuration. In the remaining of this chapter these two systems will be referred to as either "*Keplerian*" or "*Galilean*" designs. Both the *Keplerian* and *Galilean* prototypes used in the following experiments are shown respectively in figures 2.5 and 2.6. The *Galilean* compact CTIS setup offers high spatial and spectral resolution and a larger FoV of about (Fov_v $=$ FoV_h $= 9°$) but at the expense of vignetting

TABLE 2.1: System specifications for the *Keplerian* and *Galilean* CTIS designs

System	Spatial Resolution				Spectral Resolution			Dataset	
	Sensor (MP)	0^{th} order	Ground Truth	SR factor	range (nm)	step (nm)	# bands	# samples	Train/test split (%)
Keplerian	1.2	140×140	840×840	$\times 6$	$455 - 695$	10	25	496	80/20
Galilean	13	312×420	936×1260	$\times 3$	$470 - 700$	7	33	1324	92/8

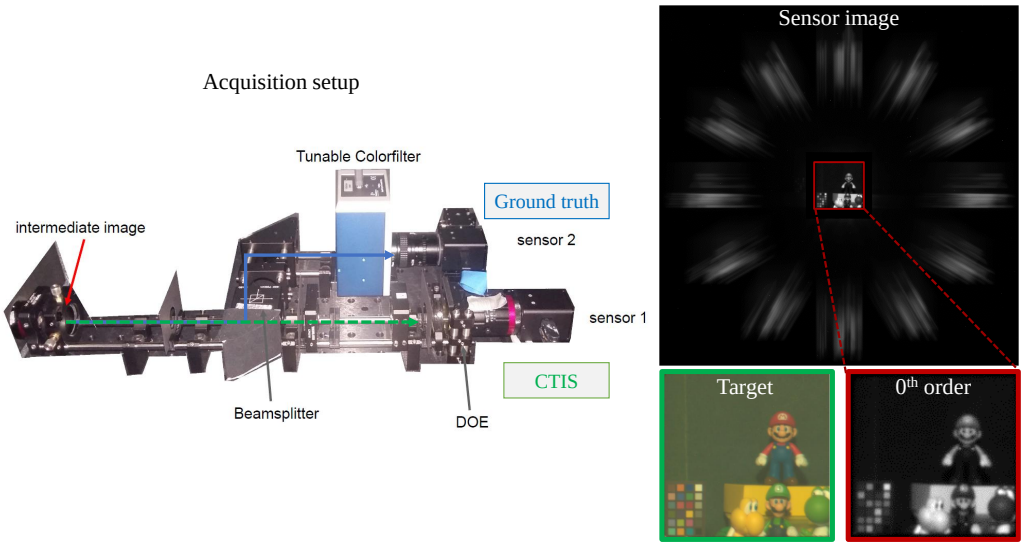


FIGURE 2.5: Data acquisition setup featuring a full-frame CTIS prototype and a ground truth camera with a Varispec™ tunable color filter (left). A sample captured CTIS measurement with a ground truth hyper-spectral image in sRGB space (right).

and image distortion affecting the 0^{th} order image as well as higher order projections which the network needs to correct for.

Data Acquisition: The full-frame *Keplerian* CTIS data acquisition setup is shown in Figure 2.5 along with an actual measurement sample and the corresponding ground truth object cube shown in sRGB space. Light is simultaneously captured by the CTIS camera and the ground truth camera using a beam splitter. 25 spectral bands are recorded for each scene, the recorded bands span the range from 455 nm to 695 nm with a spectral resolution of 10 nm in the visible range. A dataset was captured using this setup containing 495 measurements along with corresponding object cubes to train and test the proposed learning-based approaches. The DOE design produces 12 higher order diffractions surrounding the 0^{th} order image in a circular pattern.

On the other hand, figure 2.6 shows a real capture along with the acquisition setup using the *Galilean* CTIS prototype to acquire real data to train and test HSRN+ with a miniaturized CTIS camera [Ama+23a]. Ground truth data for the two setups are obtained via a secondary optical path that includes a VariSpec™ tunable color

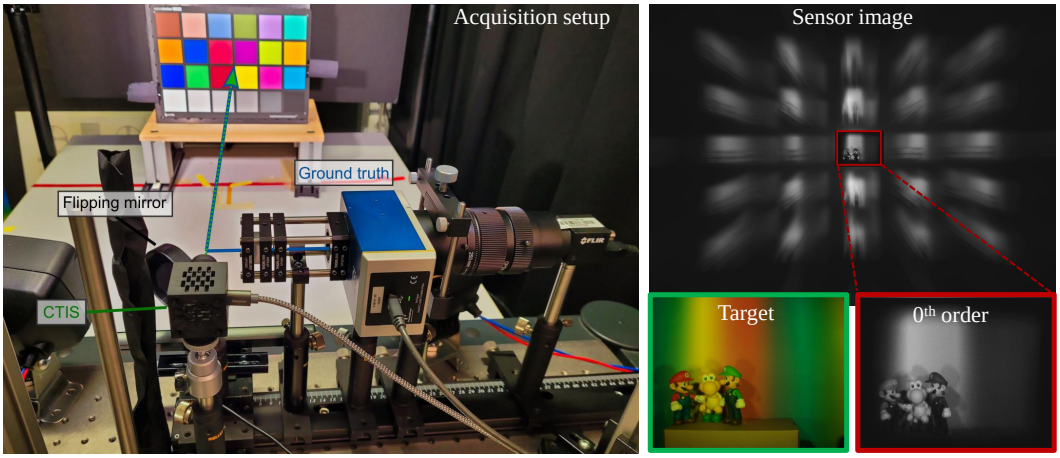


FIGURE 2.6: Data acquisition setup featuring a compact CTIS prototype and a ground truth camera with a Varispec™ tunable color filter (left). A sample captured CTIS measurement with a ground truth hyper-spectral image in sRGB space (right).

filter and a monochrome image sensor. A flipping mirror is used to redirect light to either the CTIS camera or the ground truth camera.

A large number (1.2k) of hyper-spectral images along with their corresponding CTIS measurements were captured using this setup. Part of this larger dataset captured with the *Galilean* prototype has been made public (`https://github.com/LTTM/HSIRS`) featuring 33 spectral bands spanning the spectral range from 470 nm to 700 nm with a spectral step of 7 nm and it has 592 hyper-spectral images, notice that such ground truth object cubes can be used to simulate different spectrometers and not only limited to CTIS and validate their performance accordingly. Further in depth discussion about this dataset will be presented later on in Section 2.5.5. Table 2.1 summarizes the characteristics of the two collected datasets, both of them feature scenes with varying spectral and spatial complexity including different background colors and textures with high spectral variety and multiple metameric objects and fake/real food items, a few ground truth sample images from the *Keplerian* as well as the *Galilean* setups are shown in figures 2.7 and 2.8.

2.3.4 CTIS Image Pre-Processing

The CTIS sensor image in two-dimensional image space is first cropped and reshaped into multiple three-dimensional pseudo-spectral cubes as shown in figure 2.9, this is motivated by the fact that spectral information is smeared within higher-order projections.

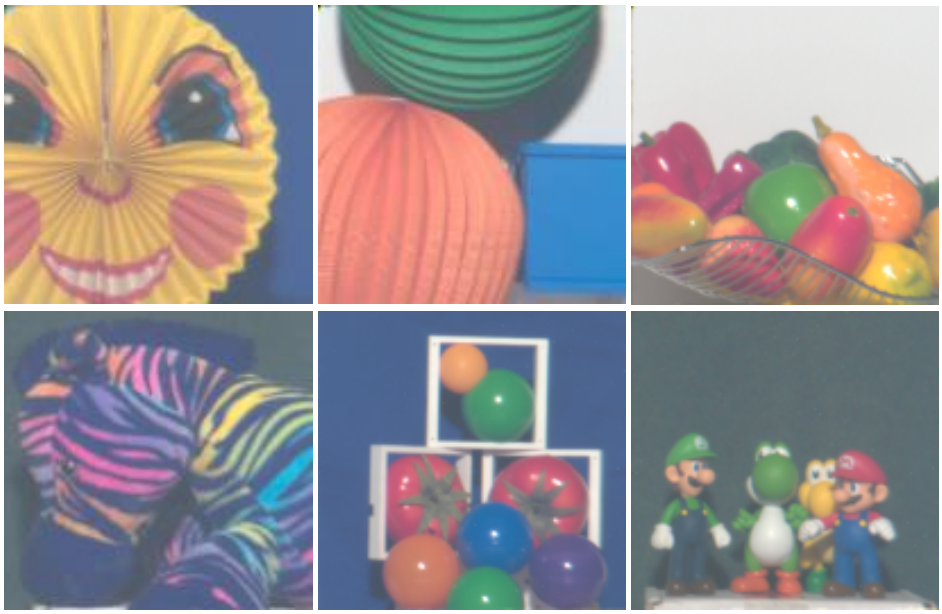


FIGURE 2.7: Sample captured ground truth images (in sRGB space) used to train and test the network on the *Keplerian* CTIS data.

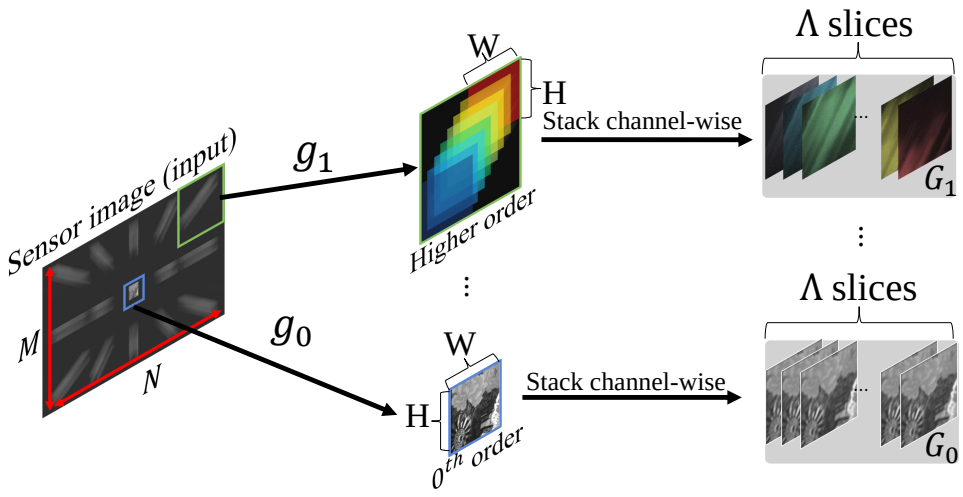


FIGURE 2.9: Cropping and reshaping of the input CTIS sensor measurement for later processing by the neural network.

The reshaping is done so that classic convolution filters can connect areas where the corresponding spectral information is distributed. Furthermore, the reshaping layer produces a data structure suitable for subsequent processing via three-dimensional convolution layers and for preserving only valid sensor regions containing actual information. In particular, given a set of P projections $\{g_p\}_{p=0}^{P-1}$ (including the 0^{th} order image with $p = 0$), let G_p denote the reshaped three-dimensional cube from g_p were Λ slices of size $H \times W$ are cropped via a sliding window (see figure 2.9) and stacked channel-wise ending up with a three-dimensional volume of

FIGURE 2.8: Sample captured ground truth images (in sRGB space) used to train and test the network on the *Galilean* CTIS data.

shape $H \times W \times \Lambda$. Notice that each channel of the cube G_p contains the latent spectral band to be reconstructed multiplexed with adjacent bands. The rechaped data volume G_0 is obtained by repeating the 0^{th} order image across the spectral dimension.

2.3.5 Learned Back Projection

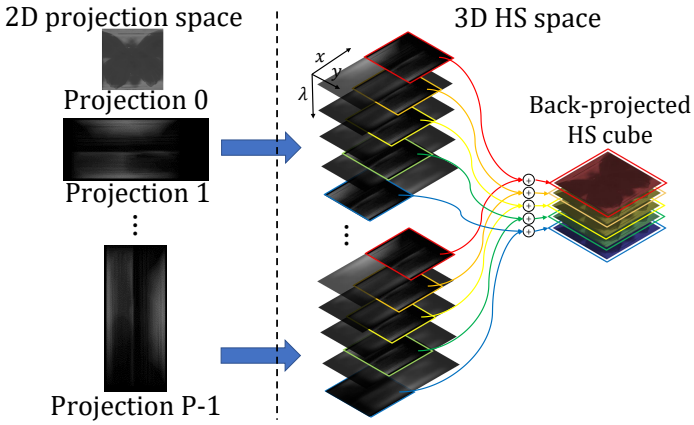


FIGURE 2.10: Back-projection of multiple 2D CTIS projections into a 3D HS cube.

Let f^{BP} denote the back-projected object cube computed from the sensor measurement g as:

$$f^{BP} = \mathbf{H}^T g \tag{2.5}$$

Where $\mathbf{H}^T$ is the adjoint operator of $\mathbf{H}$. It is easy to see that each pixel in f^{BP} is just the sum of all its contributing counterparts from g as each element of $\mathbf{H}^T$ maps

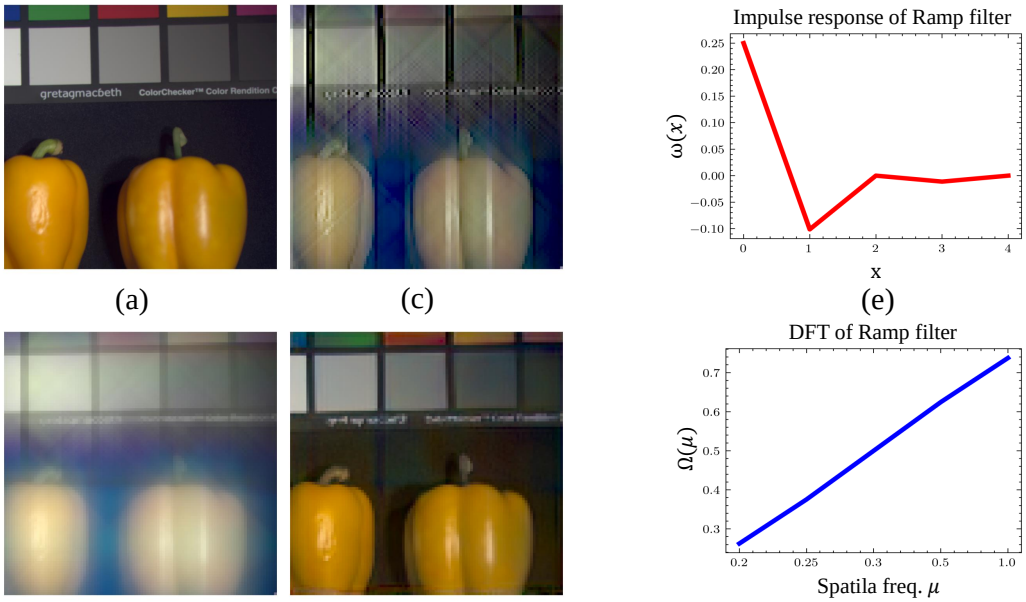


FIGURE 2.11: (a-d) are respectively the ground truth spatially super-resolved object cube, back projected image f^{BP}, filtered back projected image f^{FBP}, and object cube f^{LBP} obtained by the proposed LBP layer. (e-f) are the spatial and frequency responses of the ramp filter used to obtain f^{FBP}.

pixel contributions from the sensor back to object space. This operation, illustrated by figure 2.10, maps a low dimensional 2D projection into higher dimensional 3D hyper-spectral space and it can be implemented simply by repeating each channel from $\{G_p\}_{p=0}^{P-1}$ across the spectral dimension and summing them with equivalent spectral slices from all other projections for each wavelength to produce a first coarse rendition of the latent three-dimensional object cube:

$$f^{BP}(\lambda) = \sum_{p=0}^{P-1} G_p(\lambda) \quad \forall \quad \lambda \in [1, ..., \Lambda] \tag{2.6}$$

f^{BP} contains coarse spatio-spectral information of the latent cube but it is heavily blurred with undesirable silhouette artifacts (an example is shown in figure 2.11). To overcome the blur issue, the Filtered Back-Projection (FBP) algorithm is typically used in computed tomography scans where prior to the back-projection step, the measurements are filtered using a band limited high pass Ramp filter ω introduced by [RL71]. Such filter has the frequency response shown in figure 2.11 (f) and preserves only high frequency components in the input measurement thus reducing the amount of blur in the final back-projected image as illustrated in figure 2.11 (c). Rewriting (2.6) to account for the filtering operation results in the following expression:

$$f^{FBP}(\lambda) = \sum_{p=0}^{P-1} \omega * G_p(\lambda) \quad \forall \quad \lambda \in [1, ..., \Lambda] \tag{2.7}$$

where the symbol $*$ represents a two-dimensional convolution. Notice that ω is a

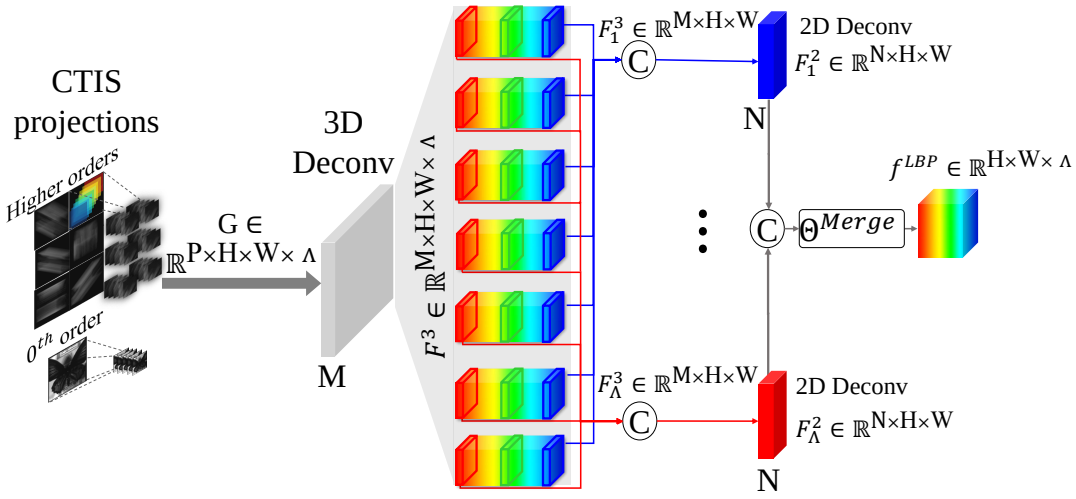


FIGURE 2.12: Schematics and workflow of the LBP layer. Such module learns to reconstruct back-projected hyper-spectral cubes in an end-to-end fashion.

fixed kernel that mainly enhances the contrast within each projection to reduce the amount of blur. However, it introduces high frequency noise and ringing artifacts due to the shape of the filter's spatial response as shown in figure 2.11 (e). The back-projection is also a global operation and it evenly maps two-dimensional projected data back into hyper-spectral space through the summation in Eq.(2.7) and it does not take into account different contributions from each higher order projection, i.e., the amount of dispersion differs for each projection leading to varying degrees of overlap between consecutive spectral bands thus higher order projections that are more dispersed carry more reliable spectral information than those dispersed across smaller sensor area. In this work, a Learned Back Projection (LBP) layer is proposed to address the aforementioned limitations. The layer architecture is illustrated in figure 2.12, LBP tunes such filtering operation to each individual projection by learning different kernel weights for each projection in an end-to-end fashion. In more detail, intra-projection correlations are learned by means of a single three-dimensional deconvolution layer [Zei+10] that takes as input the four dimensional volume cube $G \in \mathbb{R}^{P\times H\times W\times \Lambda}$ with P channels corresponding to each $\{g_p\}_{p=0}^{P-1}$. This deconvolution layer has M three-dimensional filters and produces a feature map $F^3 \in \mathbb{R}^{M\times H\times W\times \Lambda}$. Inter-projection correlations are learned via multiple parallel two-dimensional deconvolution layers: for each spectral band, all sub-feature maps from F^3 carrying distinct and complementary spatial and spectral information belonging to the same spectral band λ are concatenated channel-wise to form a 3D cube $\{F^3_\lambda\}_{\lambda=1}^{\Lambda} \in \mathbb{R}^{M\times H\times W}$ and fed into parallel two-dimensional deconvolution layers with N filters each. Lastly, the output feature maps from the previous stage are concatenated channel-wise to form $F^2 \in \mathbb{R}^{\Lambda\cdot N\times H\times W}$ and fed into a small U-net-like network $\Theta^{Merge}(\cdot)$ with multiple two-dimensional convolution layers that acts like a merger network which produces a coarse rendition of the latent HS cube through explicit supervision.

2.3.6 3D Pixel Reshuffling for Image Super-Resolution

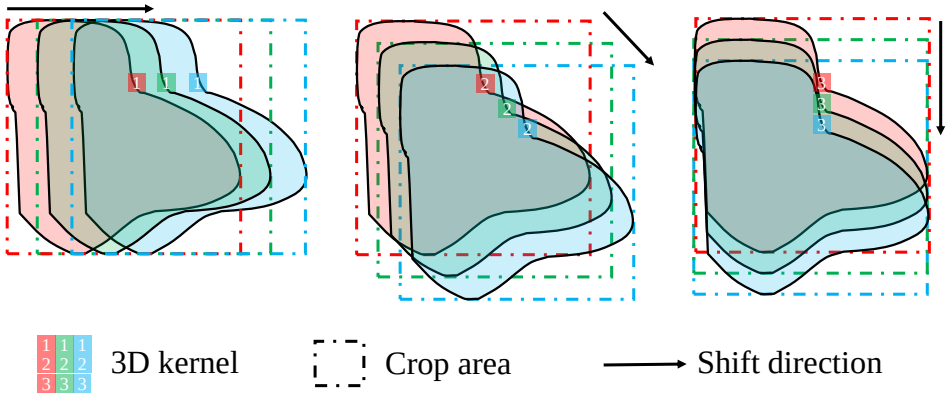


FIGURE 2.13: Simplified schematic of the three-dimensional filter's field of view taking into account higher order projections cropping and channel-wise stacking.

Local PSFs generated by the DOE carry aliased information as they differ slightly for each wavelength and for each higher order projection: each projection can be seen as a unique view of the smeared latent object cube. However, severe spatio-spectral overlap within the CTIS sensor image makes it extremely challenging to exploit such information in the context of standard multi-frame image super-resolution where the provided input data is a set of clear images of the same scene shifted with respect to each other with sub-pixel accuracy, thus spatial super-resolution becomes a straightforward operation that rely on robust image registration and merging. Nevertheless, in the case of CTIS, one might observe that sub-pixel displacements are detectable across image edges and that the smearing direction for each projection preserves image edges along that said direction, e.g., vertical edges are preserved in vertical higher order projections and so on (see figure 2.13). Given such observations, the image super-resolution sub-problem is treated in a residual learning context where a three-dimensional Sub-Pixel Convolution layer (3D-SPC) is proposed to restore high spatial frequency image features learned from distinct information provided by different higher order projections along with the 0^{th} order image, which are then combined with the coarse spatio-spectral cube generated by LBP via simple summation. Ideally, a large enough number of projections is needed to accurately restore finer spatial features of the latent cube across more gradient directions but the finite number of projections provides limited aliased information which further motivates the use of deep-learning based approaches in this case. The proposed module is inspired by spatio-temporal processing in video super-resolution [Li+19]. three-dimensional deconvolutions [Zei+10] are used to learn meaningful correlations between spatial information scattered across multiple higher diffraction orders along with the 0^{th} order image. By using multiple layers, it is possible to learn complex

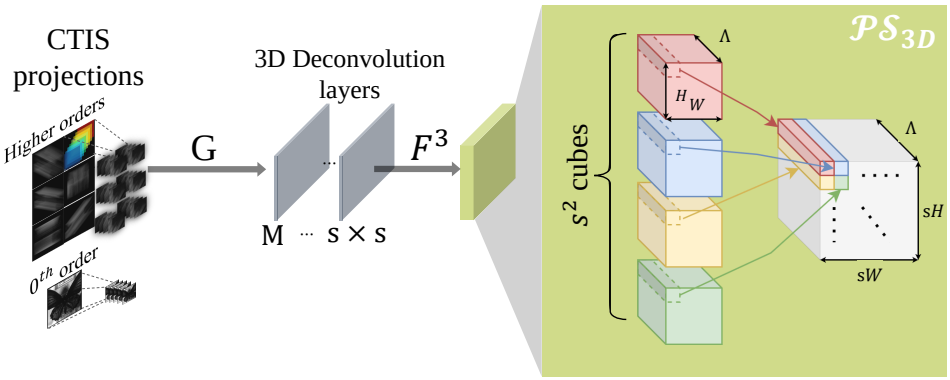


FIGURE 2.14: 3D-SPC module with sub-pixel shuffling layer. Spatio-spectral correlation from different projections are learned through deconvolution layers.

high level spatial feature maps. To this end, an adaptation of ESPCN [Shi+16] is proposed where it was originally introduced for single image super-resolution. In such approach convolution layers are applied in low resolution image space, hence it is computationally efficient and at the same time it achieves competitive image restoration results. This work's adaptation extends ESPCN in 3D space dubbed ($\mathcal{PS}_{3D}$) and performs 3D periodic pixel reshuffling with a super-resolution factor s:

$$\mathcal{PS}_{3D}(F^3)\Big(x,y,\lambda\Big) = F^3\Big[s\cdot mod(y,s) + mod(x,s) + 1, \lfloor x/s\rfloor, \lfloor y/s\rfloor, \lambda\Big] \quad \forall\lambda \in [1,...,\Lambda] \tag{2.8}$$

where $F^3 \in \mathbb{R}^{s^2\times H\times W\times\Lambda}$ is a four-dimensional feature map obtained from $G \in \mathbb{R}^{P\times H\times W\times\Lambda}$ by applying multiple three-dimensional deconvolution layers with the last layer having $s\times s$ output channels. Eq. (2.8) is illustrated by figure 2.14 and implies that for a given spectral band λ, $s\times s$ high-resolution pixels are obtained by periodically shuffling low-resolution pixels from $s\times s$ feature cubes. As illustrated by the simplified schematics in figure 2.13 each filter's receptive field "sees" different signal projections of the same latent object cube region each containing aliased pixel information and distinct spatio-spectral cues depending on the smearing direction. Mathematically such convolution can be expressed as $Out = w_3 \star \{G_p\}_{p=0}^{P-1}$, where p is the projection index, w_3 is a three-dimensional filter, $G_p \in \mathbb{R}^{H\times W\times\Lambda}$ is the reshaped tensor from a given projection g_p:

$$g_p = \mathbf{D}\mathbf{H}_p\mathbf{W}_p f_{HR} \tag{2.9}$$

where $\mathbf{D}$ is a down-sampling operator, $\mathbf{H}_p$ is a dispersion matrix that models the DOE effects, $\mathbf{W}_p$ is an affine warping matrix for sub-pixel displacement, and f_{HR} is the latent super-resolved object cube to be reconstructed. Such mathematical formulation of the local image formation model for each projection makes it easy to deduce the relationship with standard multi-frame image super-resolution approaches

where in this case $\mathbf{H}_p$ is added to account for the DOE effects. Notice that each kernel of w_3 is applied on $\{G_p\}_{p=0}^{P-1}$ with each channel carrying distinct yet complimentary spatial information.

2.4 HSRN: End-to-End Learning for CTIS

In this section HSRN is introduced. Model architecture and workflow will be discussed followed by data and training details, experimental results discussion and ablation studies. Finally, concluding remarks, observations, and future outlook concerning this approach will be presented.

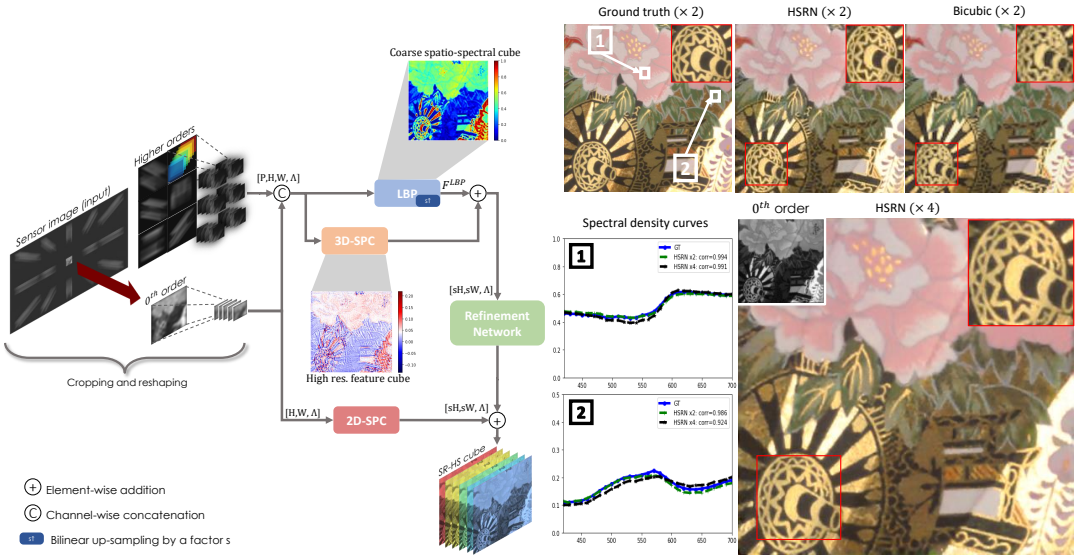


FIGURE 2.15: Proposed network architecture of HSRN (left). Sample reconstructed object cube in sRGB space and spectral density curves (right).

2.4.1 Workflow

The network architecture of HSRN is shown in figure 2.15 (left). Reshaped higher diffraction orders along with the 0^{th} order image are fed into LBP and 3D-SPC blocks simultaneously, the reason behind such approach is to exploit raw aliased pixel information directly from the image sensor for 3D-SPC without induced alterations. It is worth noting that HSRN uses a simplified and lighter version of the LBP layer described in Section 2.3.5 where the number of output channels in this case is set to $N = 1$ (see figure 2.12) and the merging network is set to be the identity operator $\Theta^{Merge} = \mathbf{I}$ that maps directly the concatenated outputs of the parallel convolution layers since there are Λ of them already. The output of LBP is up-sampled bi-linearly by a factor s to generate coarse spatio-spectral information and added up to the 3D-SPC residual output with high frequency spatial information to form an intermediate rendition of the latent object cube, this estimate is later refined using a small convolutional neural network with 7 convolution layers each with 64 filters. No

down-sampling operations, e.g., strided convolution or pooling, are used within the network in order to preserve spatial details and since the network is relatively shallow, computational burden is insignificant. The output is summed up with a super-resolved 0^{th} order obtained using the original [Shi+16] 2D-SPC method. Notice that even without such residual connection the network output will not be heavily affected (refer to ablation studies Section 2.4.5 for more details). Rather, we observed that such connection introduces robustness to noise and leads to more stable training with faster convergence on noisy data in accordance with [Zha+17]. Figure 2.15 (right) shows a reconstructed hyper-spectral image with $\times 2$ and $\times 4$ the resolution of the 0^{th} diffraction order shown in sRGB space – the conversion from hyper-spectral space to standard RGB is done via the conversion function of the CIE 1931 norm [Écl31] – and compared with a bi-cubic up-sampled reference object cube, notice that fine spatial details are restored with minimal artifacts. Spectral density curves are also shown along with the Pearson correlation coefficient between the predicted and ground truth curves.

2.4.2 Data and Training Setup

Synthetic datasets: The performance of HSRN is evaluated on synthetic CTIS data generated from three publicly available datasets: TokyoTech-31 [Mon+15], CAVE [Yas+10], and ICVL [ABS16]. A train/test split was chosen so that $\sim 75\% - 80\%$ of the total number of images are used to train the network and the rest for testing, table 2.2 summarizes the characteristics of the three different datasets used in this work.

- **TokyoTech**: Downloaded from `http://www.ok.sc.e.titech.ac.jp/res/MSI/MSIdata31.html` and contains object spectral response. All 35 scenes provided by the authors have been used to simulate CTIS images.
- **CAVE**: Downloaded from `https://www.cs.columbia.edu/CAVE/databases/multispectral/` and contains object reflectance taken in an indoor setting.
- **ICVL**: Downloaded all data samples from `https://github.com/icvl/shred?tab=readme-ov-file` and contains object spectral response of mostly outdoor scenes.

Results on a fourth dataset, Hyper-spectral Video introduced by [MH12], were used to assess real-time reconstruction performance of a pre-trained HSRN model and will be discussed in details later on. CTIS measurements are simulated with 200 spectral bands spanning the range from 420 nm to 720 nm for TokyoTech-31 and 400 nm to 700 nm for CAVE and ICVL using Fourier optics as described in Section 2.3.2. In particular, a ground truth hyper-spectral cube interpolated across the spectral dimension is convolved with a wavelength-dependent PSF to generate a CTIS sensor image with 14 higher diffraction orders (see figure 2.15). In case of noisy inputs shot noise was also introduced simulating a quantum full well capacity of 1000 photons.

TABLE 2.2: Synthetic dataset statistics and train/test split sizes.

Dataset	Spectral range (nm)	Spectral bands	Environment	Reflectance	Resolution	Size	Train split	Test split
CAVE [Yas+10]	400 - 700	31	Indoors	✓	512×512	32	25	6
ICVL [ABS16]	400 - 700	31	Outdoors	✗	1392×1300	200	160	40
TokyoTech [Mon+15]	420 - 720	31	Indoors	✗	Multi res.	35	26	9

For each scene, image regions used to simulate a sensor measurement are obtained by cropping, via sliding window of size 100×100, sub-images from the original high-resolution object cubes to simulate CTIS images in case the network is trained without spatial super-resolution, i.e., $s = 1$. Otherwise image regions are cropped with size $100s \times 100s$ (s being the up-sampling factor) effectively decreasing the sizes of each train/test split. The training data is augmented using random rotation and flipping of the ground truth object cubes spatial-wise before simulating the sensor image.

Implementation Details of Competitors: In the following, reconstruction performance of the proposed model is compared against those from other state-of-the-art approaches which were reproduced with some necessary modification to account for the shape of the input measurement:

- **Zimmermann et al. [Zim+22]:** The first cropping and reshaping layer of P2Cube has been modified to take into account the new DOE design which imprints 14 higher diffraction orders arranged in a 3×5 pattern. The rest of the network is kept as is and all hyper-parameters were also unchanged. The network was trained according to the procedure described in the original paper.

- **Ahlebaek et al. [Ahl+22]:** In the original work each higher diffraction order is cropped along with the 0^{th} order into individual sub-images each containing a whole projection then stacked channel-wise and the resulting data volume is then fed to a U-Net architecture. However, the input dimensions used in the original paper were small (250×250 pixels), instead in this work some diffraction orders are dispersed across larger sensor areas (up to 450×450 pixels) prompting the padding of the 0^{th} order (which originally has a resolution of 100×100 pixels) by 350×350 pixels. In the modified network architecture of [Ahl+22] the third downstream convolution block with 256 filters is removed and a convolution block with 256 filters in the upstream direction was added, a cropping layer after the last convolution layer was also added to restore the original spatial resolution of the 0^{th} diffraction order. Notice that the network was primarily modified only to account for the new input data dimension preserving at the same time its performance for fair comparison.

- **Expectation Maximization:** A different implementation than the one provided by [Ahl+22] was used in this work: The new implementation is

GPU-accelerated and is much faster exploiting the shift invariance property of the CTIS system matrix.

Training details: For all experimental setups, except otherwise specified, the network is trained using the Adam optimizer [KB14] with a learning rate of $1e^{-4}$ and for 500 epochs with the following loss function:

$$\mathcal{L}(I_s^\star, I_s) = MSE(I_s^\star, I_s) + \gamma \cdot MAE(I_s^\star, I_s) \tag{2.10}$$

Where $I_s^\star$ is the reconstructed object cube with spatial resolution $sH \times sW \times \Lambda$, I_s is the ground truth object cube with the same resolution. The additional use of the MAE loss term is motivated by the fact that it better preserves high spatial frequencies, these two loss terms are balanced using a discrepancy parameter γ empirically set to 0.1. A third MSE loss term is incorporated in order to force LBP to produce coarse spatio-spectral images, it is evaluated between the output of LBP $I_{LBP}^\star$ and the s-fold down-sampled reference object cube I_{s_0} to match the 0^{th} order resolution.

$$\mathcal{L}_{LBP}(I_{LBP}^\star, I_{s_0}) = MSE(I_{LBP}^\star, I_{s_0}) \tag{2.11}$$

2.4.3 Experimental Results on Synthetic Data

First, spectral reconstruction performance is evaluated without the spatial super-resolution task. Then, the network generalization capability is validated via cross dataset validation where it is trained on some dataset and tested on a completely different one. Later, results are reported of HSRN trained to jointly perform the tasks of hyper-spectral image reconstruction and spatial super-resolution with $\times 2$ and $\times 4$ increase of the resolution of the 0^{th} diffraction order image.

Hyper-spectral image reconstruction: Quantitative results on object cubes of size $100 \times 100 \times 31$ pixels reconstructed by HSRN are reported in table 2.3 along with qualitative results in figure 2.16. Quantitative and qualitative performance is compared to that of [Zim+22] and [Ahl+22]. Since both competing approaches did not tackle the problem of spatial super-resolution, the scale factor s has been set to 1, thus reducing the sub-pixel shift layer in HSRN to an identity mapping. The proposed model is able to outperform both competing approaches on all three datasets with a smaller number of trainable parameters and faster reconstruction speed at inference time. More in detail, [Zim+22] is capable of outperforming [Ahl+22] with a much lighter model size, but is in turn outperformed by HSRN that is also lighter. Figure 2.16 shows three reconstructed hyper-spectral images from TokyoTech-31, CAVE, and ICVL converted to sRGB space. HSRN is able to produce higher-quality spatial and spectral distributions with less artifacts such as color leakage and blurring.

TABLE 2.3: Quantitative comparison on multiple spectral benchmark with competing approaches.

Method	#Params (M)	Time (s) (CNN/EM)	TokyoTech-31 RMSE↓	TokyoTech-31 PSNR↑	TokyoTech-31 SSIM↑	CAVE RMSE↓	CAVE PSNR↑	CAVE SSIM↑	ICVL RMSE↓	ICVL PSNR↑	ICVL SSIM↑
Ahlebaek et al. [Ahl+22]	26.6	0.05 / ≥ 10	0.035	28.849	0.872	0.039	28.708	0.823	0.021	33.896	0.881
P2Cube [Zim+22]	1.5	0.017 / -	0.028	33.033	0.917	0.024	34.448	0.941	0.005	47.497	0.991
HSRN	**0.9**	0.010 / -	**0.025**	**33.809**	**0.941**	**0.018**	**37.282**	**0.964**	**0.004**	**48.470**	**0.995**

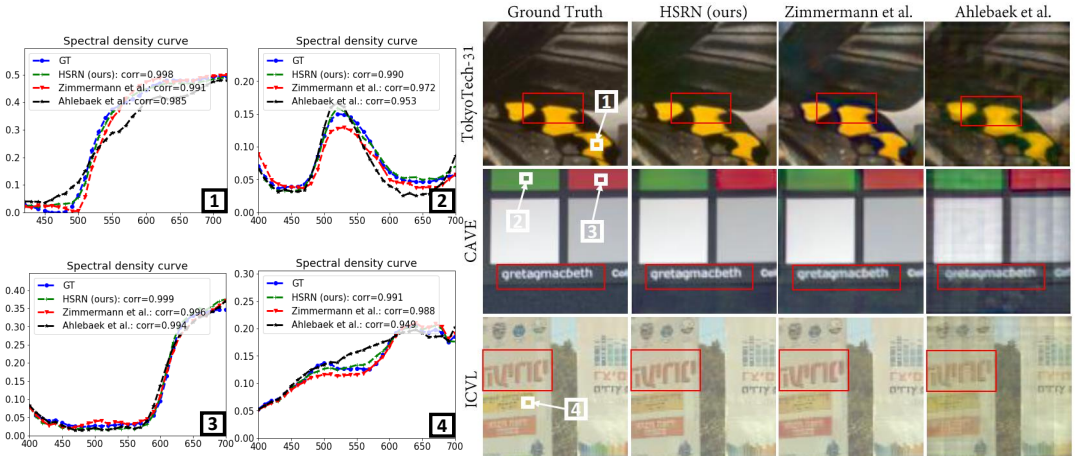


FIGURE 2.16: Reconstruction results on three different benchmarks with object cubes of size 100×100×31 pixels.

Joint spectral & spatial super-resolution reconstruction: HSRN is able to reconstruct object cubes with a ×2 and ×4 increase in spatial resolution with respect to that of the 0^{th} diffraction order image achieving a resolution suitable for most sensing applications. Quantitative performance results are shown in table 2.4 along with reconstructed samples in figures 2.17a and 2.17b. In order to validate the spatial super-resolution capability of HSRN, its performance is compared with two straightforward sequential approaches where HSRN reconstruction stage (with $s = 1$) is used and followed by either: (i) a bicubic up-sampling with refinement through multiple convolution layers trained separately or (ii) by the original ESPCN network from [Shi+16]. In the easier case of ×2 super-resolution factor, both sequential approaches are able to achieve satisfactory performance but still significantly lower than the one achieved by HSRN and with ×4 they fall short of achieving acceptable results while HSRN preserves high PSNR scores (up by roughly 7 dB on ICVL compared to [Shi+16]). Notice that the reconstruction speed on an NVIDIA RTX A6000 GPU of a 400 × 400 × 31 object cube is about 0.033 seconds. Closeup inspection of

TABLE 2.4: Hyper-spectral reconstruction and spatial super-resolution results.

Data	Scale	HSRN RMSE↓	HSRN PSNR↑	HSRN SSIM↑	Bicubic+CNN RMSE↓	Bicubic+CNN PSNR↑	Bicubic+CNN SSIM↑	Shi et al. [Shi+16] RMSE↓	Shi et al. [Shi+16] PSNR↑	Shi et al. [Shi+16] SSIM↑
TokyoTech-31	×2	**0.026**	**34.738**	**0.945**	0.033	33.495	0.914	0.029	33.030	0.928
CAVE	×2	**0.018**	**37.244**	**0.956**	0.024	35.313	0.942	0.022	35.538	0.944
ICVL	×2	**0.011**	**42.065**	**0.972**	0.025	39.371	0.958	0.018	40.623	0.965
TokyoT. + CAVE	×4	**0.033**	**32.731**	**0.907**	0.078	24.556	0.844	0.057	27.375	0.888
ICVL	×4	**0.012**	**39.661**	**0.955**	0.061	29.065	0.889	0.036	32.732	0.902

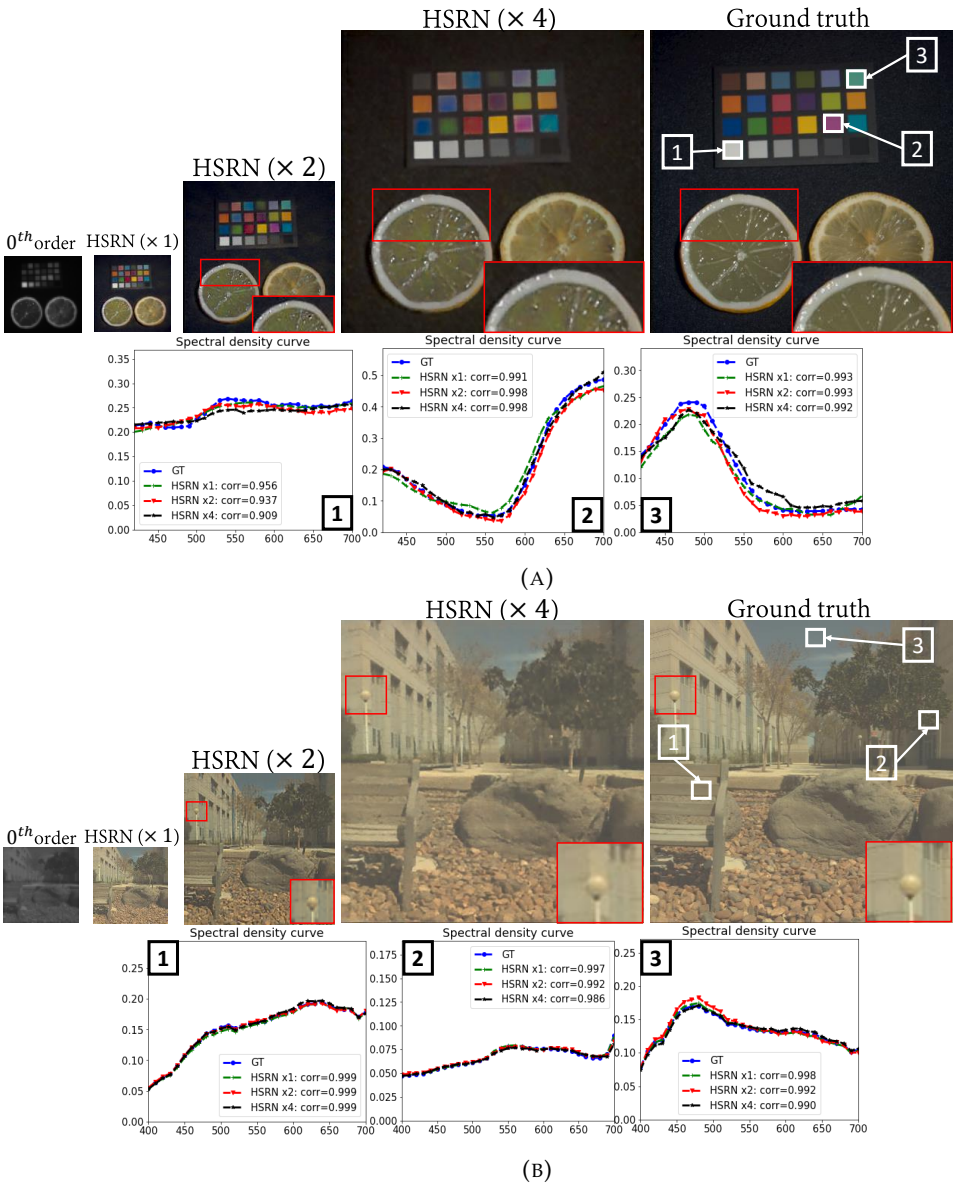


FIGURE 2.17: Hyper-spectral and super-resolution image reconstruction results of a test scene from CAVE (A) and ICVL (B).

figures 2.17a and 2.17b show that in the case of $\times 4$ super-resolution fine spatial details are restored such as the ones within the structure of the lemon in the zoomed-in image region, spectral density curves also show little deviation from the ground truth spectral data with high correlation values.

Cross dataset validation: Model transferability and generalization capability is also assessed using cross-data validation on TokyoTech-31 and CAVE by training the network on one dataset "*Source*" and testing it on the other "*Target*" so that performance is validated in both directions (*Source* $\leftrightarrows$ *Target*). Results are compared to those from [Zim+22] since it is the second best performing approach. Both architectures are trained to reconstruct object cubes with 29 spectral bands (420 $\rightarrow$ 700

TABLE 2.5: Cross-dataset validation results (↑/↓ percentages in blue).

	Scale	TokyoTech-31 → CAVE			CAVE → TokyoTech-31		
		RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑
P2Cube [Zim+22]	×1	0.025 (↑13.6%)	33.539 (↓4.5%)	0.917 (↓3.2%)	0.058 (↑114.8%)	29.931 (↓10.7%)	0.895 (↓2.9%)
HSRN	×1	0.022 (↑4.5%)	35.164 (↓1.6%)	0.948 (↓2.1%)	0.034 (↑47.8%)	31.052 (↓10.2%)	0.918 (↓2.6%)
HSRN	×2	0.022 (↑37.5%)	34.912 (↓8.4%)	0.930 (↓3.4%)	0.033 (↑50%)	31.687 (↓13%)	0.922 (↓3%)

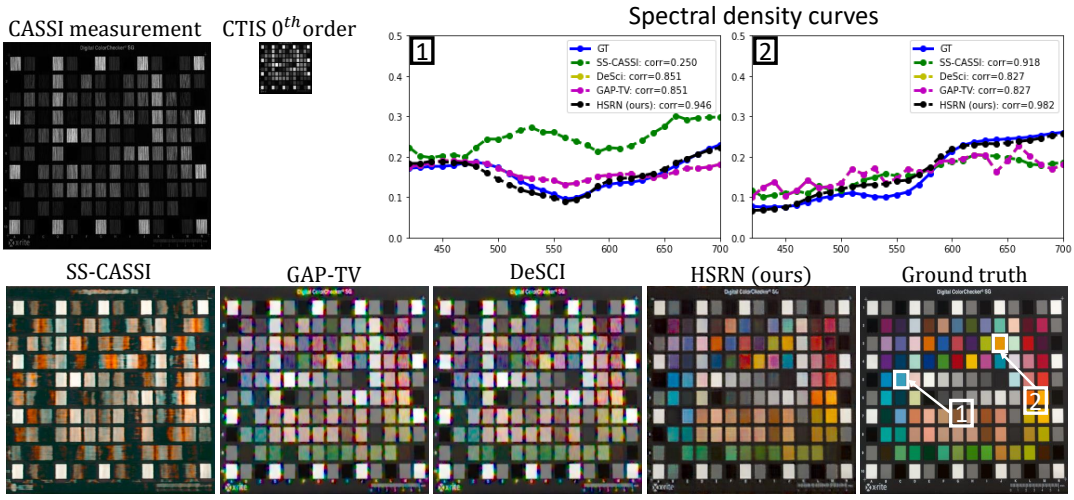
TABLE 2.6: Comparison with CASSI-based reconstruction approaches.

Method	Time	`Checkerboard`			`Butterfly`		
	(CPU-s)	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑
SS-CASSI [Men+21]	16911	0.08	18.536	0.611	0.025	29.322	0.799
GAP-TV [Yua16]	17	0.055	21.975	0.700	0.027	27.446	0.884
DeSCI [Liu+18]	4465	0.055	21.975	0.700	0.019	29.191	0.909
HSRN	0.1	**0.022**	**29.186**	**0.898**	**0.010**	**35.944**	**0.956**

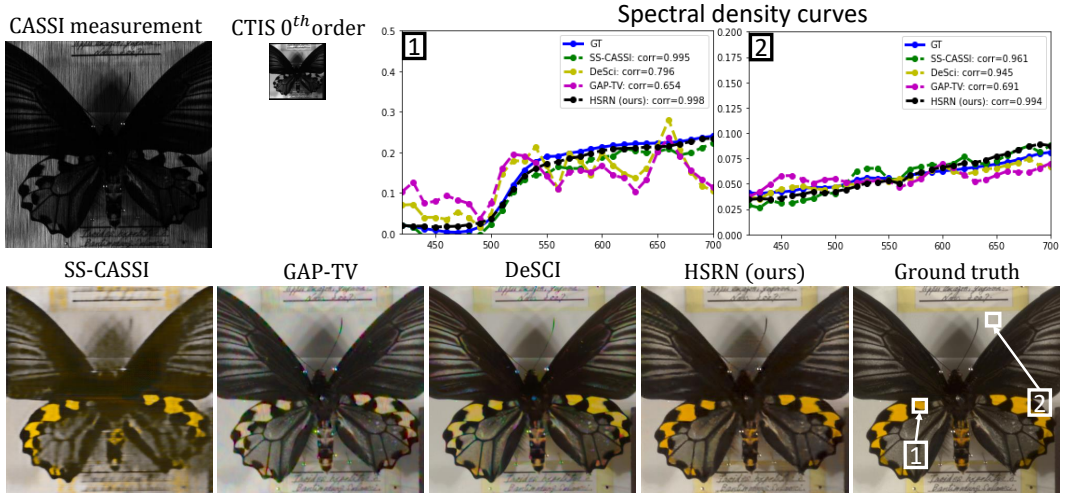
nm) with the same spatial resolution of the 0^{th} order, i.e., $s = 1$. Furthermore, results of HSRN with spatial super-resolution capability with a factor of $\times 2$ are also reported. All evaluation metrics shown in table 2.5 prove the generalization capability of the proposed model where it achieves better performance with respect to [Zim+22] at the base resolution and maintains good performance with $\times 2$ spatial super-resolution on the *Target* test set.

Comparison with CASSI-based reconstruction methods: To showcase the suitability of CTIS systems for spectral image sensing beyond the spatial resolution limitation of the 0^{th} order image, reconstruction performance comparison with other spectral reconstruction methods designed for CASSI systems is presented using two different test images (`Checkerboard` and `Butterfly`) picked from the test set of TokyoTech-31. The reconstructed object cubes are of size $400 \times 400 \times 29$ pixels. Quantitative reconstruction performance reported in table 2.6 compares HSRN, that performs joint spectral reconstruction and $\times 4$ image super-resolution, with model-based iterative approaches for CASSI, that optimize directly on a super-resolved measurement coded by an aperture mask which degrades the spatial resolution of the measurement. Despite the low resolution of the 0^{th} order image, HSRN is able to reconstruct object cubes with higher spatial and spectral accuracy achieving a gain of 6 to 8 dB of PSNR compared to CASSI reconstruction methods. Visual results in figures 2.18a and 2.18b confirm the numerical metrics showing how HSNR restores very fine spatial details on the `Checkerboard` and `Butterfly` images with minimal color artifacts due to inaccuracies in the reconstructed spectrum.

Hyper-spectral video reconstruction: In this experiment HSRN is pre-trained using a combination of TokyoTech-31 and CAVE datasets for the task of joint spectral reconstruction and spatial super-resolution with $\times 4$ that of the 0^{th} diffraction order image. The trained model is then tested on the Hyper-Spectral Video dataset [MH12] mimicking real-time performance using an NVIDIA RTX A6000 GPU. This dataset contains 31 video frames each with a spatial resolution of 752×480 and



(A)



(B)

FIGURE 2.18: (A) Reconstruction of the `checkerboard` test target. (B) reconstruction of the `butterfly` test target.

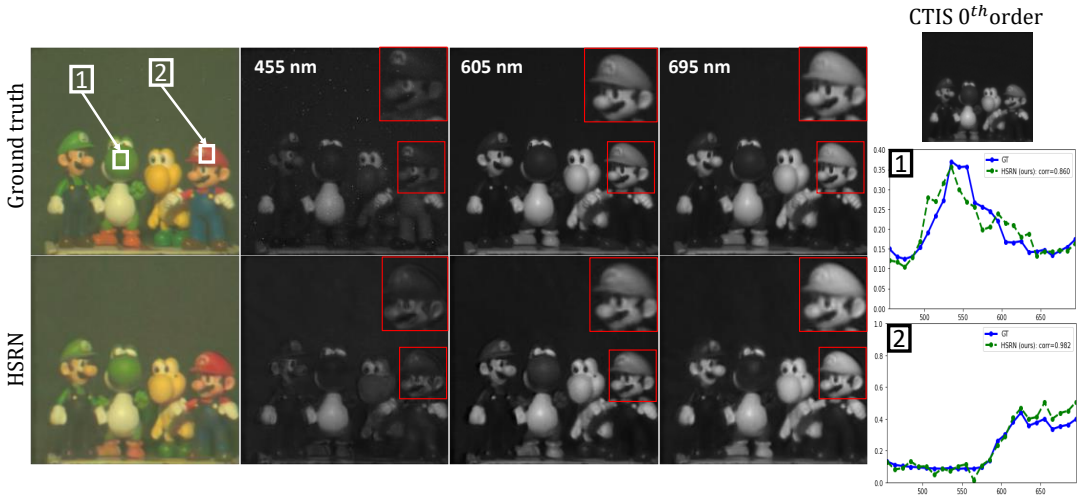


FIGURE 2.19: Sample of a reconstructed image from CTIS real data captured with the *Keplerian* setup along with spectral density curves of some selected regions.

33 spectral bands ($400 - 720$ nm), only 29 spectral bands are used in accordance with the number of bands used in training data and the object cubes are spatially resized to 100×100 pixels in order to simulate CTIS images. The reconstruction time of a cube of size $400 \times 400 \times 29$ is roughly 0.033 seconds. Real-time reconstruction performance is shown in the publicly available video clips at: `https://medialab.dei.unipd.it/paper_data/HSRN_CTIS/` at the original frame rate of 30 fps and slowed down to 2 fps for better visualization. HSRN is able to reconstruct fine object details with good spectral accuracy specially considering that it was trained on different data.

2.4.4 Experimental Results on Real Data

The real CTIS dataset captured by the full-frame *Keplerian* CTIS prototype is used to train and test HSRN in the real domain. The network reconstructs hyper-spectral images with spatial resolution of 278×278 pixels, that is $\times 2$ the resolution of the 0^{th} diffraction order and with 25 spectral bands spanning the range from 455 nm to 695 nm. A real sample of a reconstructed cube is shown in figure 2.19 in sRGB space together with three individual spectral bands (455 nm, 605 nm, and 695 nm) and selected spectral density curves of two different image regions. Achieving good reconstruction quality on real data is challenging due to the limitation imposed by the misalignment between CTIS sensor measurement and the ground truth object cubes since the acquisition setup has two optical paths and therefore misalignment is inevitable. HSRN+ is later proposed to tackle this issue.

TABLE 2.7: Quantitative results of different ablation experiments.

LBP	3D-SPC	Residual	TokyoTech-31 (w/ shot noise) PSNR↑
✗	✓	✓	30.214
✓	✗	✓	30.521
✓	✓	✗	31.501
✓	✓	✓	**31.832**

2.4.5 Ablation Studies

The contribution of each module in HSRN is evaluated by testing the network performance without said module. The network was trained for 250 epochs in all ablation experiments and tested on the TokyoTech-31 test set with shot noise. Table 2.7 shows how each component gives a relevant and non-overlapping contribution to the results. Worth noting that the residual connection from 2D-SPC does not lead to significant degradation when excluded from the model architecture while 3D-SPC and LBP have bigger influence on the overall performance of the network.

2.4.6 Concluding Remarks

The above work introduced a joint approach for hyper-spectral image reconstruction and spatial super-resolution from CTIS data tackling for the first time the major shortcomings of such system and providing an efficient model capable of performing reconstructions in real-time. By exploiting side information from higher diffraction orders HSRN was able to produce object cubes with fine spatial details and up to $\times 4$ the spatial resolution of the 0^{th} diffraction order image. That being said, the main limitations of this approach are two-fold: *Spectral-wise*, small angles of parallel projection, i.e., the amount the HS cube is smeared in a given projection, may hinder the reconstruction quality as spectral bands severally overlap each other at the sensor and the network struggles to accurately resolve them. *Spatial-wise*, enough higher order projections are needed to reach acceptable reconstruction accuracy specially for large super-resolution factors, e.g.,$\times 4$, as more complementary information would be available which in turns require larger sensor area. Furthermore, performance in real data lags behind simulations due to the large domain gap, to this end, a new enhanced model will be presented which focuses on improving the real CTIS camera and acquisition setup and the reconstruction performance on real data captured by such system.

2.5 HSRN+: Multi-Scale Learning for CTIS

In this section HSRN+ will be presented. Model architecture and workflow will be discussed followed by data and training details and then results discussion and ablation experiments. Finally, concluding remarks, observations, and future outlook will be discussed.

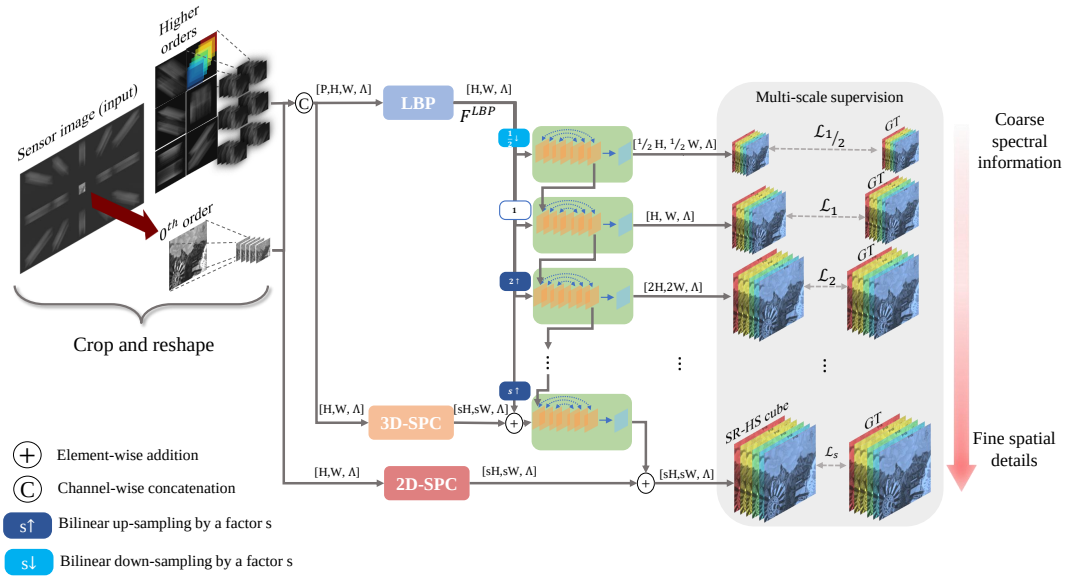


FIGURE 2.20: HSRN+ architecture: Coarse rendition of the latent object cube is obtained by LBP output added to image features with high spatial frequencies restored by 3D-SPC module. The output is built incrementally from coarse to fine scales each supervised with a dedicated loss function.

2.5.1 Workflow

HSRN+ preserves the building blocks of HSRN and adds to that a multi-scale learning-based architecture. The full network architecture of HSRN+ is shown in figure 2.20. Higher diffraction orders along with the 0^{th} order image are first reshaped into three dimentional cubes and then fed into LBP and 3D-SPC modules. Worth noting that HSRN+ uses the original architecture of the LBP layer described in Section 2.3.5 where the number of output channels in this case is set to $N = 32$ (See figure 2.12) and the merging network $\Theta^{Merge}(\cdot)$ contains seven consecutive convolution layers and takes as input a dense feature volume with $\Lambda \times N$ channels obtained by concatenating the outputs of the previous parallel convolution layers since there are Λ of them. $\Theta^{Merge}(\cdot)$ produces a coarse spatio-spectral data volume which is summed up to the 3D-SPC residual output with high frequency spatial information to form an intermediate object cube, the estimate is later refined using a multi-scale refinement strategy.

2.5.2 Multi-Scale Supervision and Training Details

In HSRN+, the three-dimensional object cube is built incrementally in a coarse-to-fine fashion as shown in figure 2.20. The idea is to build coarse spectral information first then gradually refine the reconstruction quality spatial-wise in each subsequent stage where coarse knowledge from previous levels is propagated to the next ones to gradually ease the reconstruction burden as the resolution scale increases. Coarse feature maps from the second to last convolution layer in each refinement network are up-sampled and concatenated with the re-sampled LBP output to form the input

to the next finer stage. The refinement networks (shown in the green boxes) used to produce the final output at each level have the same U-net-like architecture with 8 convolution layers and skip connections in between. Multi-level loss functions are therefore used to train the model, each dedicated to supervise the output at different spatial resolutions. The loss corresponding to a given super-resolution factor s is used to back-propagate gradients across the network. In each level, the spatial resolution of the output image is doubled, starting with half the original resolution of the 0^{th} order image, the total number of stages depends on the target resolution factor s. Therefore, the loss function used to train the network on synthetic data for a given factor s is:

$$\mathcal{L}_s^{Synth}(I_s^\star, I_s) = MSE(I_s^\star, I_s) + \gamma \cdot MAE(I_s^\star, I_s) \tag{2.12}$$

Where $I_s^\star$ is the reconstructed object cube with spatial resolution $sH \times sW \times \Lambda$, I_s is the ground truth object cube with the same resolution. The use of the MAE component is motivated by the fact that it is better at preserving high spatial frequencies. The discrepancy parameter γ is set to 0.1 in the following experiments. LBP is also explicitly supervised with a separate loss term:

$$\mathcal{L}_{LBP}(I_{LBP}^\star, I_{s_0}) = MSE(I_{LBP}^\star, I_{s_0}) \tag{2.13}$$

Where $I_{LBP}^\star$ is the LBP output and I_{s_0} is the ground truth object cube down-sampled to match the 0^{th} order resolution.

In the case where the network is trained using real data and due to the inherent misalignment and slight perspective shift between the ground truth image and the CTIS sensor image (see Section 2.5.4 for more details) an additional loss term is further incorporated, more specifically, a contextual loss [MTZM18] is added which is agnostic to pixel location and only measures the similarity between feature vectors extracted, using a pre-trained VGG model [SZ14], from the reconstructed object cube with the ones extracted from ground truth spectral images. The contribution of such loss term is demonstrated in figure 2.21 where sharper image edges with better reconstruction quality are obtained in contrast to using the standard pixel-wise loss terms only. The new loss function used to train the network on real data is therefore:

$$\mathcal{L}_s^{Real}(I_s^\star, I_s) = MSE(I_s^\star, I_s) + \gamma \cdot MAE(I_s^\star, I_s) + \mathcal{L}_{CX}(I_s^\star, I_s) \tag{2.14}$$

Where $\mathcal{L}_{CX}$ is the contextual loss from [MTZM18].

It is worth noting that since the feature extractor VGG network was originally trained on RGB images from ImageNet [Den+09a], it expects as input a RGB image. A straightforward approach would be to convert both predicted and ground truth object cubes into sRGB space before evaluating such loss term but this operation would inevitably lead to the loss of spectral information due to the conversion process into sRGB space. A better approach is to feed the feature extractor for each input

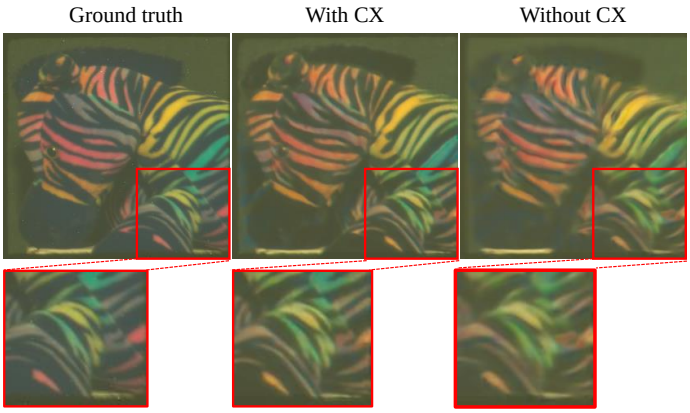


FIGURE 2.21: A reconstructed hyper-spectral image (in sRGB space) using a network trained with and without the CX loss term on real data. Notice how most of the undesirable blurring artifacts are corrected for.

training sample three randomly chosen spectral channels from the prediction as well as the reference object cubes where the central band is always fixed at the center of the spectrum, i.e., green, and the other two bands are randomly selected from the two extreme regions of the spectrum. In this way one can ensure that spectral information is preserved during training all while exploiting spatial features generated by the VGG network. The same data split was used as in HSRN [MGZ22] for all experiments using synthetic data. The network is trained for 500 epochs using Adam optimizer and a learning rate of 1×10^{-4}.

2.5.3 Experimental Results on Synthetic Data

In this section reconstruction results on synthetically generated CTIS data are presented and discussed. Furthermore, qualitative and quantitative comparisons with other competing state-of-the-art approaches are presented.

Spectral reconstruction: The reconstruction performance of HSRN+ is compared to other closely related CTIS approaches. Since none of the proposed models so far tackled the issue of spatial super-resolution, at least from a computational point of view, one can start by comparing the reconstruction performance using HSRN and HSRN+ without performing super-resolution, i.e., $s = 1$, with that from other competing approaches. It is worth noting that recently Yuan et al. [Yua+23] proposed a multi-sensor fusion approach to reconstruct spatially super-resolved images

TABLE 2.8: Quantitative comparison on multiple spectral datasets with competing approaches.

Method	#Params (M)	Time (s)	TokyoTech-31			CAVE			ICVL		
		(CNN/EM)	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑
Ahlebaek et al. [Ahl+22]	26.6	0.05 / $\geq$ 10	0.035	28.849	0.872	0.039	28.708	0.823	0.021	33.896	0.881
P2Cube [Zim+22]	1.5	0.017 / -	0.028	33.033	0.917	0.024	34.448	0.941	0.005	47.497	0.991
HSRN [MGZ22]	**0.9**	**0.010** / -	0.025	33.809	0.941	0.018	37.282	0.964	0.004	48.470	0.995
HSRN+ Small	**0.9**	**0.010** / -	0.022	35.980	0.944	0.018	37.372	0.964	0.004	49.857	0.995
HSRN+	1.9	0.078 / -	**0.020**	**36.357**	**0.956**	**0.017**	**37.695**	**0.967**	**0.003**	**50.780**	**0.995**

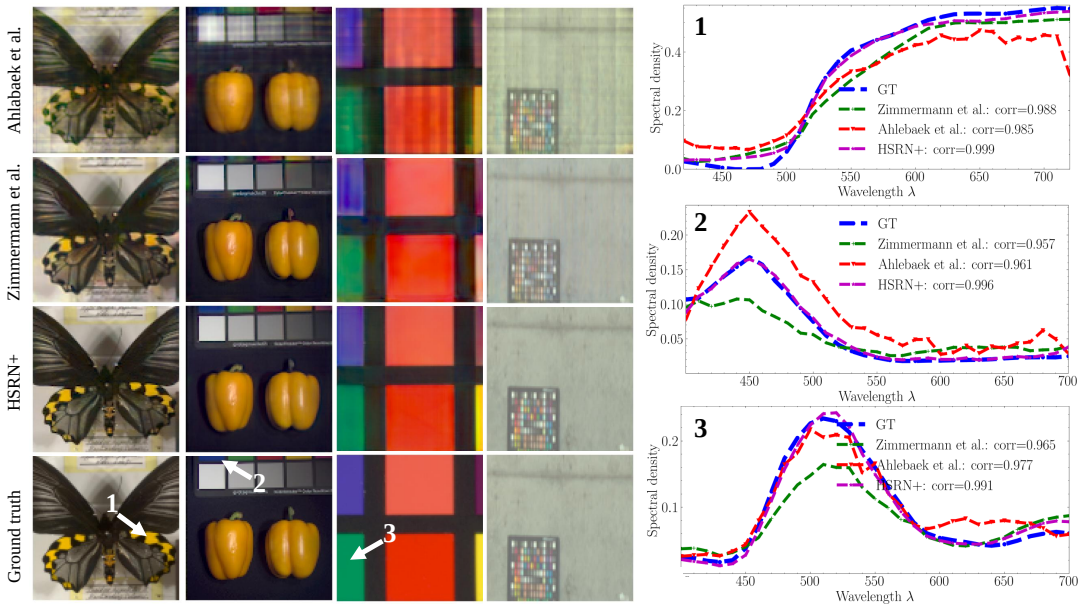


FIGURE 2.22: Reconstruction results on simulated CTIS data without spatial super-resolution: the spatial resolution of the reconstructed hyper-spectral cubes is 100×100 pixels. Spectral density distributions for some chosen regions are shown on the right along with Pearson correlation coefficient between the predicted and ground truth curves.

from CTIS measurements using an additional RGB image with high spatial resolution along with the low resolution CTIS measurement. Differently, the proposed approach in this work requires a single sensor measurement as input without the need for additional hardware. Table 2.8 and figure 2.22 show quantitative and qualitative results on the three publicly available datasets (ICVL [ABS16], TokyoTech-31 [Mon+15], and CAVE [Yas+10]) used previously to train and test HSRN. In order to highlight the improvements of HSRN+ with respect to its predecessor, regardless of network size, performance metrics for a smaller version of HSRN+ are also reported in the table 2.8. In particular, in this setting the number of convolution filters in each reconstruction network (showed in green boxes in figure 2.20) is reduced from 64 to 32 resulting in an overall size of 0.9 M trainable parameters roughly on par with those of HSRN.

The three proposed model variants shown in the gray rows of table 2.8 outperform the current state-of-the-art on all three benchmarks and across all evaluation metrics. HSRN+ achieves the best PSNR on the TokyoTech-31 dataset [Mon+15] with a gain of over 2.5 dB with respect to HSRN. Even with roughly the same number of trainable parameters, a gain of over 2 dB is achieved by HSRN+ "small" which highlights the contribution of the employed multi-scale learning strategy. The same performance gain can be observed on the ICVL dataset [ABS16] with a gain of 2.3 dB obtained by HSRN+ "small" with respect to HSRN. On the other hand, a smaller performance gain with respect to HSRN can be observed for CAVE [Yas+10], as the training set of such dataset is quite small and the network tends to over-fit it. Hence, early stopping has been used to prevent such behaviour. However, even

TABLE 2.9: Quantitative comparison for the joint tasks of spectral reconstruction and spatial super-resolution on three benchmarks.

Data	Scale	Shi et al. [Shi+16]			HSRN [MGZ22]			HSRN+		
		RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑
TokyoTech-31	×2	0.029	33.030	0.928	0.026	34.738	0.945	**0.023**	**36.452**	**0.953**
CAVE	×2	0.022	35.538	0.944	0.018	37.244	0.956	**0.018**	**37.264**	**0.958**
ICVL	×2	0.018	40.623	0.965	0.011	42.065	0.972	**0.010**	**43.119**	**0.976**
TokyoTech-31 + CAVE	×4	0.057	27.375	0.888	0.033	32.731	0.907	**0.031**	**32.969**	**0.918**
ICVL	×4	0.036	32.732	0.902	0.012	39.661	0.955	**0.010**	**41.805**	**0.967**

on the CAVE dataset, the three proposed variants outperform competitors [Ahl+22; Zim+22] by significant PSNR, SSIM, and RMSE margins. Such numerical evaluation is also reflected in the reconstructed hyper-spectral images shown in figure 2.22, where reconstructed hyper-spectral images are shown in sRGB space. HSRN+ produces sharper image details and suppresses unwanted reconstruction artifacts such as color leakages, chromatic aberrations, and unwanted blur.

Joint spectral reconstruction and image super-resolution: To assess the network's spatial super-resolution capability, HSRN+ is trained on synthetic CTIS data with spatial super-resolution factors of $s = 2$ and $s = 4$ that of the the 0^{th} order image, which has a spatial resolution of 100×100 pixels. Quantitative as well as qualitative results are shown in table 2.9 and figure 2.23. Due to the lack of sufficient number of images within each individual dataset, the training sets of CAVE and TokyoTech-31 are combined and used to train HSRN+ in the case of $\times 4$ spatial super-resolution.

Results from Shi et al. [Shi+16] in table 2.9 correspond to a sequential architecture where a low resolution object cube is first reconstructed using HSRN [MGZ22] and then individual spectral bands are spatially super-resolved using the original ESPCN approach of [Shi+16]. HSRN+ achieves the highest scores for all training settings and outperforms [Shi+16] by substantial margins simply because it leverages additional spatial information scattered across higher order projections. This observation can be further consolidated by looking at figure 2.23 where spatially super-resolved hyper-spectral images obtained from HSRN [MGZ22] as well as HSRN+ are shown in sRGB space along with the ones obtained from ESRGAN [Wan+18] and a simple bi-cubic interpolation. Notice that the super-resolution approaches highlighted in green in figure 2.23 receive as input a decimated RGB image obtained by converting the ground truth hyper-spectral cube to sRGB space and then downsampling it by a factor s so they only perform spatial super-resolution/interpolation. Whereas the input to the two proposed model variants are the CTIS sensor measurements with gray scale decimated 0^{th} diffraction order image. Nevertheless, the two variants are able to not only produce images with significantly better spatial quality but also recover the full scene spectrum.

Reconstruction with noisy CTIS data: A realistic noise model is incorporated in the CTIS sensor image simulation pipeline. Relevant noise sources in this case are

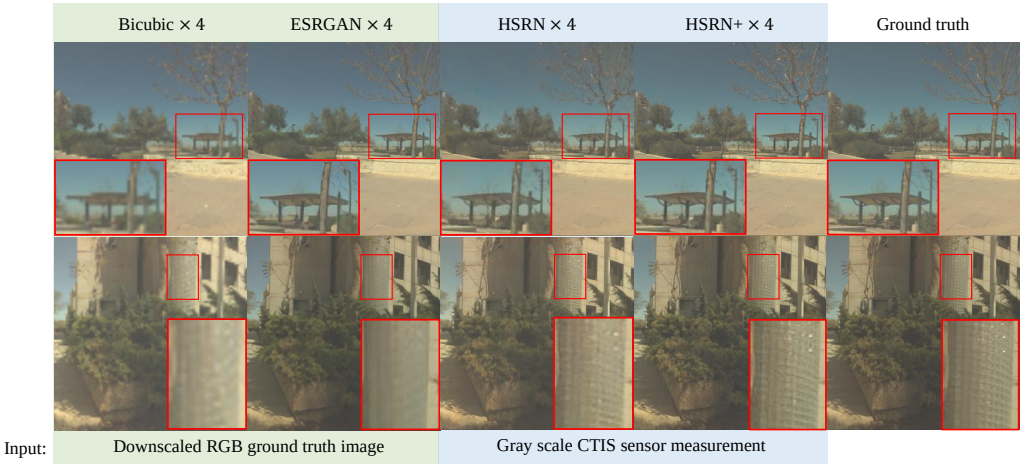


FIGURE 2.23: Spatial super-resolution performance comparison ($s = 4$) with ESRGAN [Wan+18] and a simple bi-cubic interpolation. Both HSRN and HSRN+ take as input a gray scale compressed CTIS measurement and produce spatially super-resolved spectral cubes.

read and shot noise, the former comes from inaccuracies in the sensor readout circuit and is dominant in dark image regions where signal levels are low, the latter is related to the quantum nature of light and photon arrival statistics. A realistic model should take into account the effects of both sources. Read noise can be modeled with a Gaussian distribution with zero mean while shot noise is signal dependent and is modeled by a Poisson distribution but can also be modeled as a Gaussian distribution with the variance being the pixel value. The noise model used in this pipeline is that of Foi et al. [Foi+08]:

$$\sigma(x) = \sqrt{\alpha \cdot y(x) + \beta} \tag{2.15}$$

Where σ is the pixel-dependent standard deviation of the overall noise level at pixel x, while α and β are respectively the variances of shot and read noises, and $y(x)$ is the input pixel value. In the case of CTIS sensor measurements, the values of α and β are chosen to match those estimated from the real CTIS prototype. In particular, β is estimated as in [LTO12] capturing a dark scene (where $x \approx 0$), while α can be estimated with a white input image where shot noise is dominant. Due to different light efficiencies of the DOE for the 0^{th} order image and higher order diffraction projections, two values are estimated for each variance corresponding to the 0^{th} order image (α_0, β_0) and higher order projections (α_H, β_H). The estimated values from the real CTIS camera are:

$$\begin{cases} \alpha_0 = 7.8e^{-3} & , \quad \alpha_H = 2.7e^{-3} \\ \beta_0 = 1.0e^{-4} & , \quad \beta_H = 1.1e^{-4} \end{cases} \tag{2.16}$$

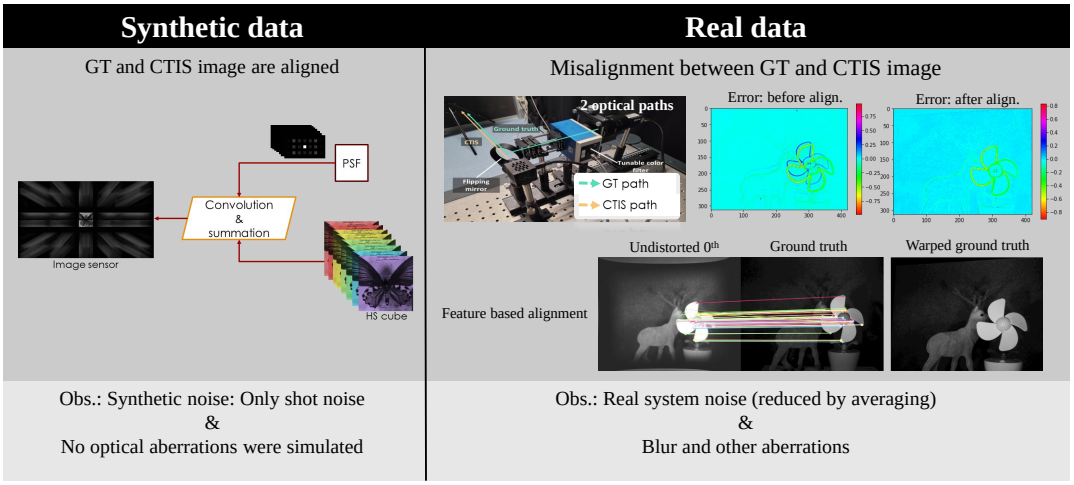


FIGURE 2.24: Simplistic data simulation pipeline (left) versus real data captured using the *Galilean* CTIS prototype (right).

It will be shown in the next section that real noise is effectively suppressed in real reconstruction results.

2.5.4 Experimental Results on Real Data

Real data misalignment: As mentioned before, there is a misalignment between the CTIS sensor measurements and ground truth object cubes which, if not accounted for appropriately, leads to sub-optimal reconstruction performance as losses are evaluated between pixels with incorrect locations. To tackle this issue, in a first step all captured images in the real CTIS datasets are aligned using feature based homography estimation and in a second step pixel location-agnostic loss terms are used in the training which was described in Section 2.5.2. Figure 2.24 illustrates the difference between the simulated CTIS data using simplistic assumptions and excluding optical aberrations, and that of real captured data with the *Galilean* prototype. The alignment step is done using the 0^{th} diffraction order image since it is undispersed and the corresponding ground truth image, i.e., a gray scale image of the object cube in sRGB space, distortion is corrected for in the 0^{th} order image using calibrated CTIS camera matrix an then features are detected in both images and matched accordingly in order to estimate the transformation between them. The estimated homography matrix is then used to warp the ground truth object cube to match the 0^{th} order image. Note that this alignment process is not optimal for every pair of images and might fail sometimes, the alignment accuracy depends on the number of correctly matched features and some scenes lack sufficient number of feature points which might leads to an ill-conditioned transformation matrix and thus wrong warping results. Figure 2.24 shows an example of matched image features from a real data sample where the alignment error is shown before and after attempting image alignment and the warped ground truth image. To tackle this issue, the contextual loss term is added during training as described before.

TABLE 2.10: Quantitative metrics achieved by HSRN and HSRN+ on real data.

Network	*Keplerian* (×6)			*Galilean* (×2)		
	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑
HSRN [MGZ22]	0.042	28.189	0.892	0.022	46.127	0.967
HSRN+	**0.031**	**30.751**	**0.916**	**0.0081**	**47.016**	**0.973**

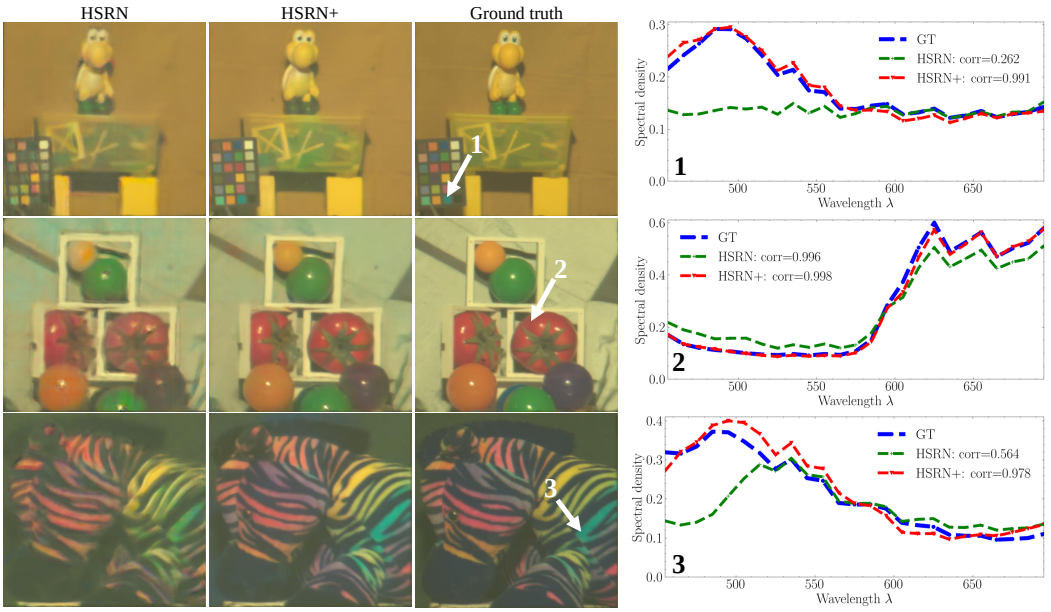


FIGURE 2.25: Reconstruction results on real data captured by the *Keplerian* setup with a super-resolution factor of ×6 that of the 0^{th} diffraction order image along with spectral density curves.

Reconstruction results: Model performance on the two real datasets described earlier and captured by the two CTIS setups are presented and discussed here. Quantitative as well as qualitative results obtained by the baselines HSRN and its subsequent variant HSRN+ are reported in table 2.10 and shown in figures 2.25 and 2.26.

The quality of the reconstruction is in-line with the reported quantitative metrics as HSRN+ is able to recover the spectral information of the scene and produce spatially super-resolved object cubes with fine spatial details even with a relatively large super-resolution factor (×6 in the case of *Keplerian* setup). Notice that the original spatial resolution of ground truth images is achieved when using a ×6 super-resolution factor.

Figure 2.26 shows individual spectral bands from a reconstructed CTIS image captured using the *Galilean* setup, in this case, due to the large size of the input CTIS image (13 MP), a super-resolution factor of only ×2 is used when training the two baselines. Notice that for smaller wavelengths the image sensor has a lower light efficiency and the noise becomes predominant in such region of the spectrum thus negatively affecting the reconstruction quality of the object cube in the small wavelength regions (close to blue).

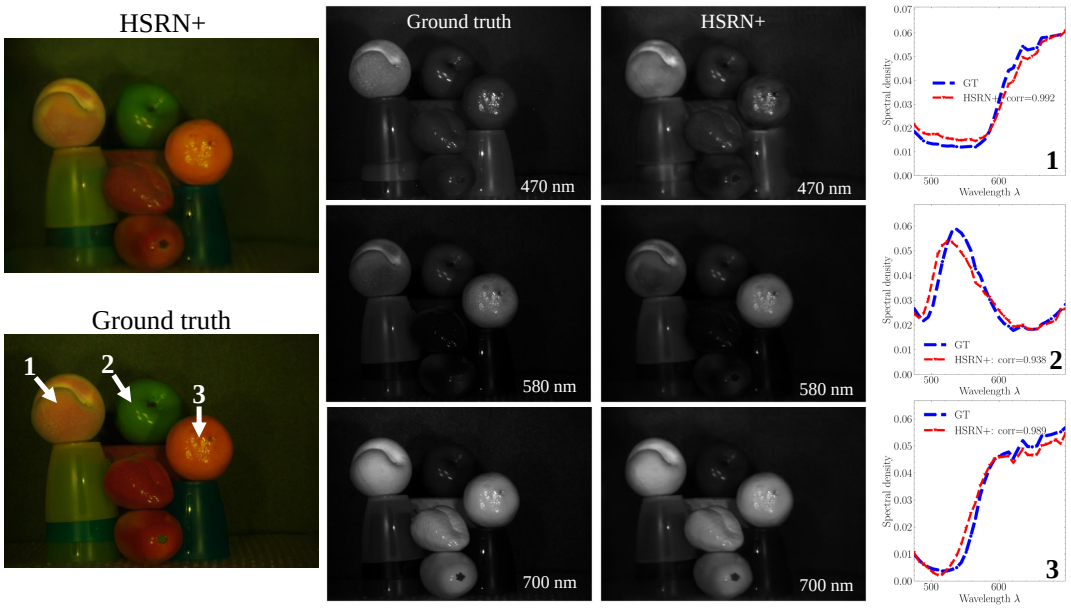


FIGURE 2.26: Reconstruction results on real data captured by the *Galilean* setup with a super-resolution factor of $\times 2$ with respect to the 0^{th} diffraction order image. Individual spectral bands are shown separately along with spectral density curves of some chosen image regions.

Comparison with the EM solver: To further the performance of HSRN+ with respect to the standard iterative EM algorithm, visual comparison results are shown in figure 2.27 to showcase performance gain both spatial- and spectral-wise where quantitative metrics such as PSNR and SSIM values are also reported for each test image. EM recovers spectral information of the 0^{th} order diffraction image and cannot, by design, handle optical aberrations such as distortion and vignetting. Furthermore, undesirable artifacts, namely halos and severe blur patterns can be observed in the reconstructed images by EM due to the lack of a sufficient number of higher order projections. On the other hand, HSRN+ is not affected by the shortcomings of the EM solver. Notably, the reconstructed hyper-spectral images do not suffer from vignetting and undesirable blur artifacts which the network preemptively corrects for owing to the large network learning capacity as it can reconstruct images with significantly better quality with limited number of projections. Figure 2.28 shows the ability to recover spectral information by HSRN+ and EM of a color patch and a real and fake (made of plastic) lemons, the proposed model is more in-line with the ground truth spectrum compared to the output of the EM solver. It is worth noting that the images reconstructed by HSRN+ tend to have the same perspective as the ground truth images, while the ones produced by the EM solver have the same perspective as the 0^{th} diffraction order image, so quantitative measures are calculated after aligning the output of the EM solver with the corresponding ground truth images.

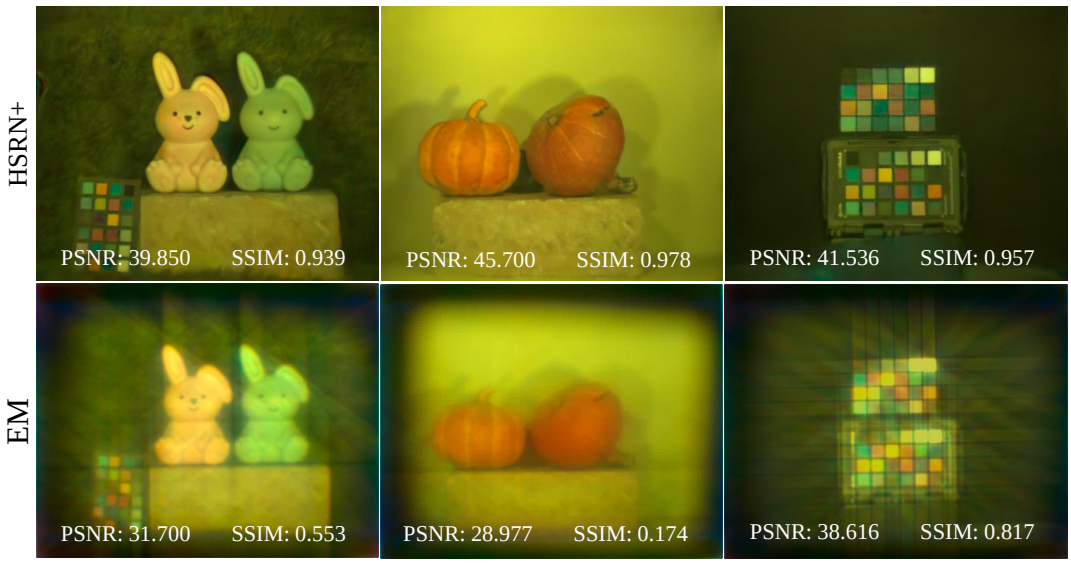


FIGURE 2.27: Reconstruction results from CTIS sensor measurements using the conventional EM solver and the proposed network HSRN+, quantitative metrics (PSNR and SSIM) are also reported.

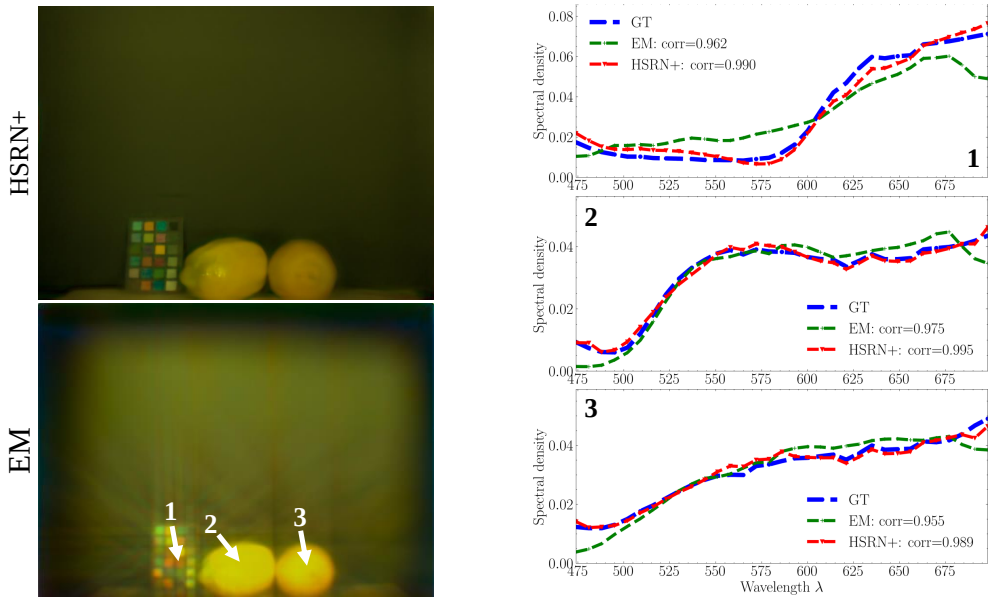
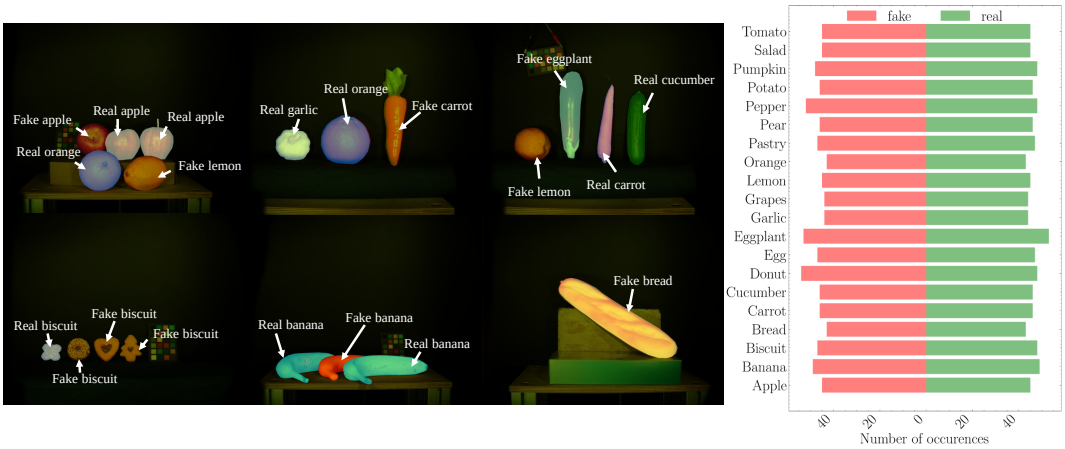


FIGURE 2.28: Recovered spectral density curves of various image regions namely a color patch, artificial lemon, and a real lemon. Spectral density curves are normalized by the are under the curve.

2.5.5 Material Characterization

Material characterization is a direct application of spectral imaging. The aim is to identify and classify different materials and compositions of materials in a scene based on their spectral signature. Such task can be formulated as a dense per pixel labeling, i.e., semantic image segmentation. However, very few datasets that deal with semantic segmentation from hyper-spectral data in indoor settings have been



(a) (b)

FIGURE 2.29: (A) Sample images with segmentation maps (super-imposed on top of the RGB image) of HSIRS containing real/fake food items. (B) Number of real/fake instances for each class in the dataset.

proposed so far, one of the few is *FVgNet* by Makarenko et al. [Mak+22]. Still, large enough high quality annotated data that enables end-to-end learning of deep neural networks is still scarce. In this work, a new dataset is introduced: HSIRS "**H**igh quality **S**pectral **I**mage **R**econstruction and **S**egmentation" dataset, a large scale dataset that contains high quality hyper-spectral images along with accurate and manually annotated segmentation maps to enable end-to-end learning for the tasks of spectral image reconstruction and semantic segmentation of different food items in indoor settings. Additionally, the ground truth images from HSIRS can be used to simulate any snapshot spectral imaging device and is therefore versatile and suitable for various spectral reconstruction and spatial image super-resolution tasks. HSIRS contains 592 hyper-spectral cubes with spatial resolution of 2048×2048 pixels. The spectral bands are in the visible spectrum range of 470 nm up to 700 nm with 7 nm spectral steps leading to a total of 33 bands. In addition, the dataset features 20 food classes as shown in figure 2.29, notice that each one of these classes have real and fake instances leading to a total of 40 distinct semantic classes. The number and layout of objects varies across scenes along with different background colors and overall scene complexity providing a good benchmark not only for semantic image segmentation but also for other downstream tasks such as spectral reconstruction and spatial image super-resolution.

As a semantic segmentation baseline, a ResU-net architecture [Dia+20] is trained to segment images from HSIRS using either hyper-spectral data or CTIS measurements directly as inputs compared to segmentation results achieved using RGB images as input (obtained by converting hyper-spectral object cubes into sRGB space

TABLE 2.11: Quantitative metrics for the semantic image segmentation task on the test set of HSIRS, all metrics are expressed in (%).

Input	mIoU	F1	Pixel Prec.	Pixel Acc.
RGB	85.62	90.50	89.90	92.05
CTIS Measurement	89.44	91.08	92.57	94.26
Hyper-spectral	**91.38**	**94.59**	**94.44**	**94.37**

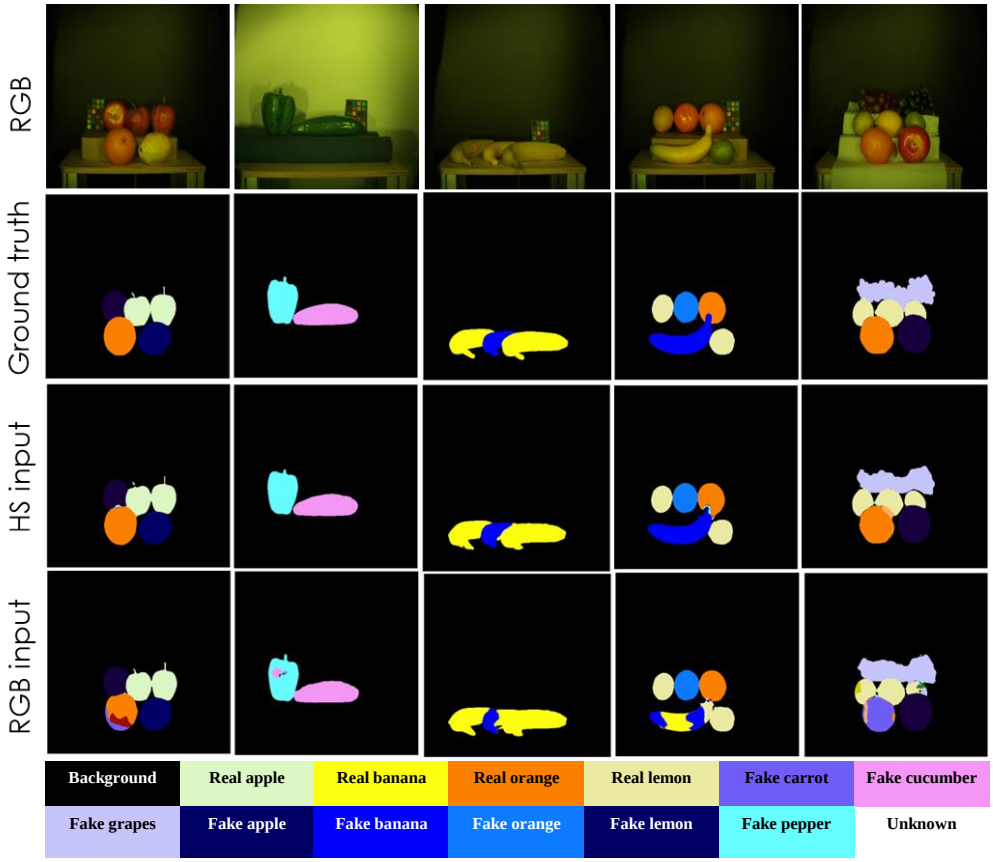


FIGURE 2.30: Sample predicted segmentation maps using either RGB or hyper-spectral data as input to the network.

using the CIE norm). Input images have been resized to 512×512 pixels. Quantitative results are shown in table 2.11 where widely used evaluation metrics for the task of semantic segmentation are reported, namely the mean Intersection over Union (mIoU), Pixel Precision, Pixel Accuracy, and F1 score. Results from table 2.11 highlight the contribution of spectral data, even with spatio-spectral multiplexing in the case of the CTIS sensor measurement as input, in achieving better segmentation performance compared to conventional RGB inputs. In particular, performances using directly the CTIS image are better than just RGB but lower than the ones obtained with the hyper-spectral data. Some predicted segmentation maps are shown in figure 2.30 where the predicted segmentaion maps using hyper-spectral data are more consistent with those of the ground truth comared to using RGB data as input to the network.

In addition, the confusion matrices for a set of chosen semantic classes are shown

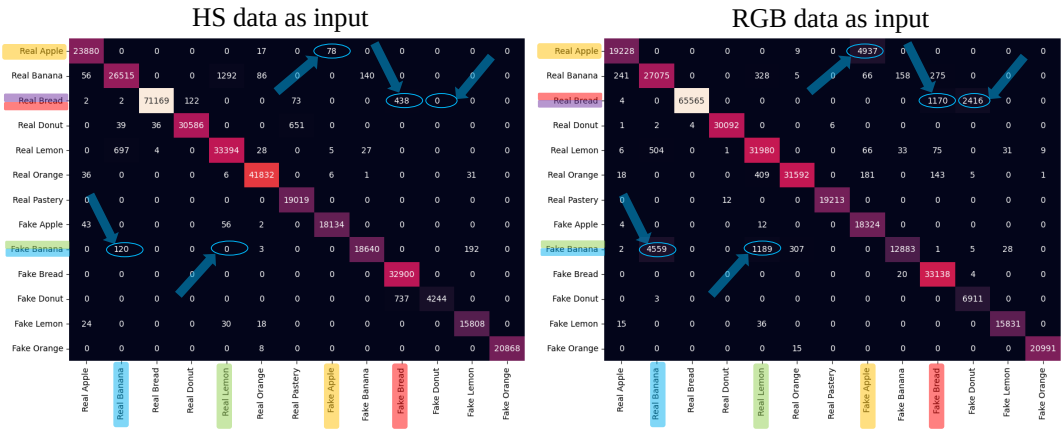


FIGURE 2.31: Confusion matrices of a subset of chosen semantic classes concerning the network trained on RGB data (left) and hyper-spectral data (right).

TABLE 2.12: Quantitative metrics for several ablations studies conducted to assess the contribution of each module within HSRN+.

LBP	3D-SPC	MSL	$\mathcal{L}_{LBP}$	TokyoTech-31 + CAVE ($\times 4$)		
				RMSE↓	PSNR↑	SSIM↑
✓	✓	✓	✓	**0.031**	**32.969**	**0.918**
✗	✓	✓	✓	0.038	31.456	0.899
✓	✗	✓	✓	0.043	30.648	0.883
✓	✓	✗	✓	0.037	31.970	0.901
✓	✓	✓	✗	0.038	31.371	0.904

in figure 2.31, highlighting further the contribution of spectral data with respect to standard RGB images: the larger spectral information content can be leveraged to enhance the segmentation performance and predict more accurate segmentation maps with better distinction between real/fake instances of the same food items.

2.5.6 Ablation Studies

The contribution of each module in HSRN+ is evaluated by testing the network performance without said module. In the following ablation experiments the network is trained and and tested on data combined from CAVE and TokyoTech-31 datasets with a spatial super-resolution factor $s = 4$. Table 2.12 provides three different evaluation metrics namely the RMSE, PSNR, and the SSIM. Quantitative results in table 2.12 highlights each component's relevant contribution.

Contribution of LBP Not only does LBP add a degree of interpretability to the whole network architecture, but it helps to achieve better spectral reconstruction quality of the final object cube as indicated by the quantitative metrics reported in the second row of table 2.12 where the PSNR decreases by 1.5 dB with respect to the baseline when LBP is omitted. Notice that in this case the number of output channels, i.e., number of filters, of the three-dimensional deconvolution layer shown

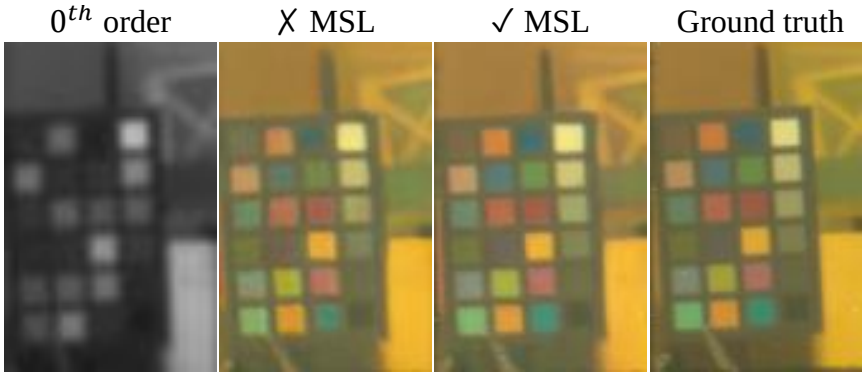


FIGURE 2.32: Color checker board reconstructed using two variants of HSRN+ with (✓) and without (✗) MSL with a spatial super-resolution factor of $\times 6$.

in figure 2.12 is set to $M = 1$ instead of $M = 32$ and a single 2D deconvolution layer is used afterwards in place of the parallel workflow of LBP.

Contribution of 3D-SPC Three-dimensional pixel reshuffling maps low-resolution feature maps into the high-resolution image space. Given the fact that such low-resolution features carry complementary spatial information from multiple higher diffraction orders that is needed to restore the high spatial resolution object cube, it is anticipated that when such information is discarded the spatial quality of the reconstructed object cubes will be degraded. In order to validate such observation HSRN+ is trained without the 3D-SPC block but keeping 2D-SPC since it performs spatial-super-resolution of the 0^{th} diffraction order image without considering spatial information scattered across higher order projections, the rest of the architecture is kept unchanged. Quantitative results from the third row table 2.12 show that all metrics dropped significantly compared to those of the baseline reported in the first row.

Contribution of Multi-Scale Learning (MSL) The use of a multi-scale learning approach eases the reconstruction burden as the spatial super-resolution factor increases and thus the object cube is reconstructed incrementally using a coarse to fine manner. It can be observed in table 2.12 that the network performance using a single output stage lags behind that of the original architecture with a PSNR drop of approximately 1 dB. Such behaviour is more prominent when using real CTIS data where the MSL strategy helps to reconstruct finer details spatial-wise in-line with ground truth data as shown in figure 2.32.

Contributions of different loss terms Further investigation is dedicated to the contributions of the LBP loss term (Eq. 2.13) in the context of synthetic data and the contextual loss term for real misaligned data. An explicit end-to-end supervision of the LBP module through $\mathcal{L}_{LBP}$ enables it to produce a coarse rendition of the latent hyper-spectral cube that is in turn up-sampled and used as part of the input

to each subsequent refinement stage as shown in figure 2.20, easing the reconstruction burden incrementally as the spatial resolution factor increases. Omitting such explicit supervision leads not only to a loss of the model's interpretability, but also a worst overall performance as reported in the last row of table 2.12.

2.5.7 Concluding Remarks

Observations: In this work HSRN+ was introduced, a joint framework for hyperspectral image reconstruction and spatial super-resolution from CTIS measurements via multi-scale refinement strategy. Reconstruction performance has been demonstrated on synthetic as well as real CTIS data with high super-resolution factors (up to $\times 6$ for the *Keplerian* CTIS design). Furthermore, a direct downstream task of spectroscopy has been presented and a large scale dataset has been proposed for hyperspectral image reconstruction and semantic image segmentation with high quality manually annotated segmentation maps which is publicly available.

Domain gap: In the case of synthetic data and in order to address the trade-off between simulation accuracy and speed during training, the CTIS Point Spread Function (PSF) was simulated based on simplistic assumptions and straightforward Fourier optics principles, which do not take into account the manufactured DOE deviation from the original Computer Generated Hologram (CGH) design and much of the optical aberrations present in real captures. Such assumptions rely primarily on the spatial shift invariance of the system's PSF. The deviation between synthetic and real measurements makes network adaptation to the real domain challenging. This was the main motivation for capturing real CTIS data. Since the spectral range and spectral resolution are different from synthetic data, the network was trained and tested on each real dataset separately (the ones captured by the *Keplerian* and the *Galilean* CTIS designs). However, since real calibrated PSFs capture all optical aberrations and fully characterize the CTIS system, it is worth investigating using real calibrated PSFs to simulate CTIS sensor measurements and use tat to train HSRN+ and later test it on real captures. A possible data simulation and and real-time testing workflow is shown in figure 2.33 where calibrated PSFs for each projection and for each wavelength are used to simulate realistic CTIS captures using already captured hyper-spectral cubes. HSRN+ can be trained on such synthetic data and later tested in real-world scenarios. [b]

2.5.8 Multi-Aperture CTIS (MACTIS)

A persistent problem encountered when using a single DOE CTIS with a single aperture is the overlap between consecutive higher diffraction orders, e.g. lower wavelength of ± 2 order start to overlap with higher wavelengths of the preceding ± 1 order and so on, which negatively affect the reconstruction qualty of spectral bands

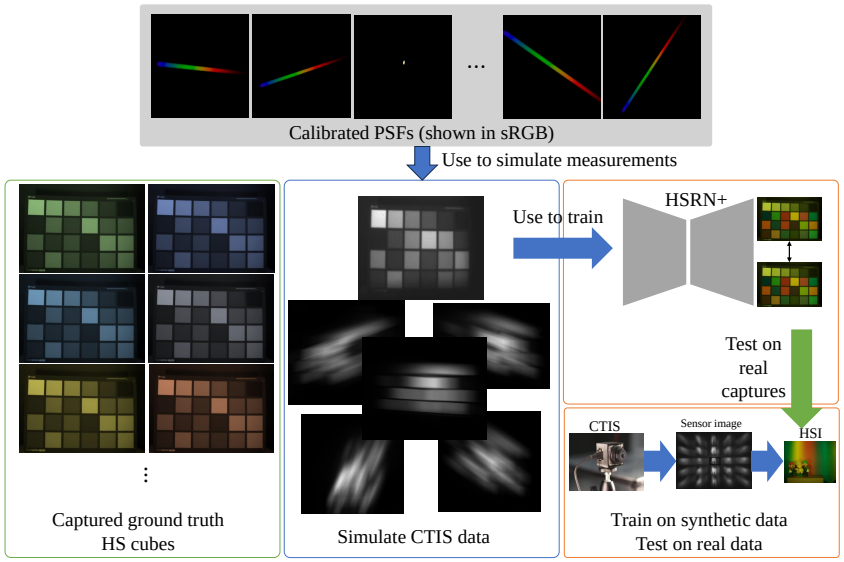
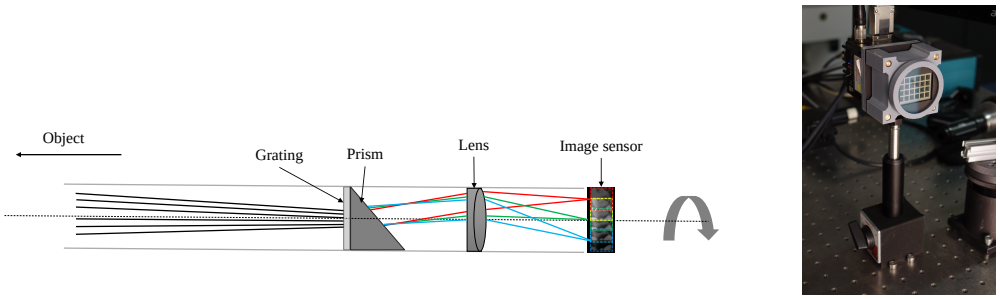


FIGURE 2.33: CTIS data simulation with calibrated PSFs and model training/testing pipeline.



(A) (B)

FIGURE 2.34: (A) Simplified schematics of a single MACTIS aperture. (B) MACTIS system prototype.

in those overlapping regions specially when using model-based reconstruction approaches. Furthermore, the system has a low light throughput and is inflexible with complex optics setup. To address those shortcomings, Amann et al. [Ama+23b] introduced a next generation CTIS system that features a multi-aperture design where several apertures placed in a grid formation each containing a diffraction grating followed by a prism (GRISM) to create only one projection per aperture as shown in the schematics in figure 2.34a . The "0^{th} order" image is obtained using a clear aperture selected among the ones in the grid. Such design is flexible as each projection angle can be obtained by rotating the GRISM of a specific aperture in a desired direction and with a desired rotation angle. MACTIS prototype is shown in figure 2.34b if has 24 aperture in a 6×4 formation, among them is a clear aperture for the undispersed image. The image sensor used in the prototype is a Ximea CMV50000 47.5 MP monochromatic sensor and the recoverable spectral range is from 450 nm to 750 nm with a 7 nm spectral resolution. The proposed reconstruction models i.e.,

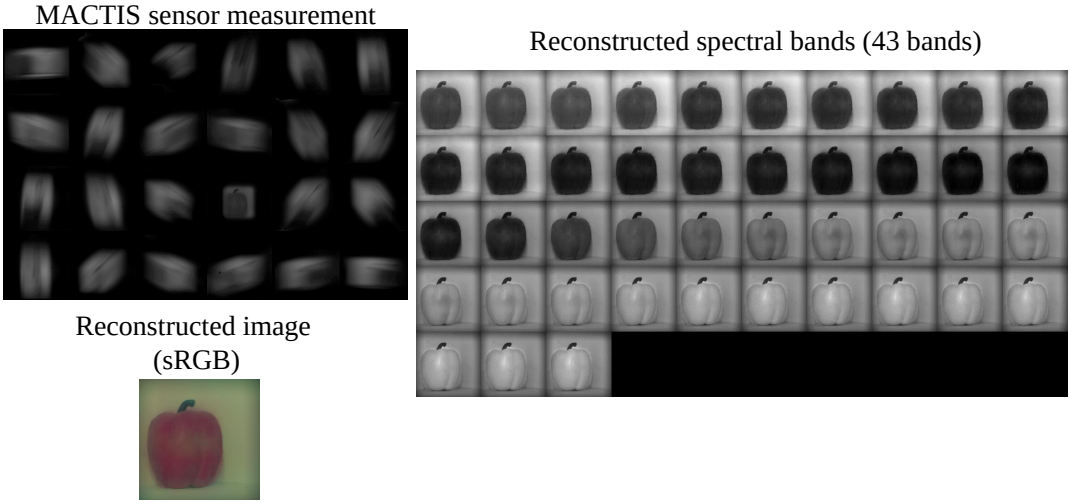


FIGURE 2.35: RAW MACTIS sensor measurement (left). Reconstructed spectral bands and sRGB image of the scene (right).

HSRN or HSRN+, are still valid in the case and can recover object cubes from MACTIS measurements. Reconstruction results of a scene containing a real red pepper using HSRN+ trained solely on synthetic data as discussed above can be seen in figure 2.35.

Chapter 3

Holographic Phase Imaging

3.1 Introduction

Dennis Gabor introduced holography in his seminal work [Gab49] where he demonstrated true three dimensional imaging capability by simultaneously recording coherent light's intensity and direction, i.e., phase, within a two dimensional interference pattern. Gabor's prototype, depicted in figure 3.1 on the right, embodies an in-line holographic imaging setup. Here, an object of interest is illuminated by a coherent light source, generating a diffraction pattern at the detector plane obtained through the interference of two waves: one scattered by the object and another unobstructed background wave passing through the object plane. This interference pattern encodes both amplitude and phase information of the wave scattered by the object, thus characterizing its complete complex transmission distribution, which can be recovered using phase retrieval-based reconstruction techniques. The ability to capture phase information allows for interesting applications such as phase imaging in microscopy [Mir+12], where it allows to generate image contrast for transparent ultra-thin microscopic specimens like single cells and other biological samples.

Advancements in phase imaging have facilitated close-up non-invasive inspection of living cells [PDP18], with applications in medicine [Par+08], biology [KVB08], and neuroscience [Cin+17]. Highly accurate interferometry, driven by holographic imaging, finds applications in high-precision engineering [VS70] and material science. Phase contrast microscopy [BS42] also offers phase imaging capability but lacks quantitative measurement ability. In contrast, holography is able to quantify exactly the amount of light phase shift making it suitable for applications requiring precise measurements, such as accurate tolerance estimations of microscopic features and three-dimensional object reconstruction [Zie+19], additionally, in-line holographic setups can be used in compact, mobile, and lensless imaging systems free of any optical aberrations. Despite its desirable features, in-line holography typically requires iterative numerical reconstruction [GS94] to retrieve high-quality images and suppress undesirable artifacts, such as the *twin image* [Zha+18], which corresponds to the latent field's complex conjugate recorded as a byproduct by the image sensor and appears as an out-of-focus, blurry image superimposed onto the true sharp one, resulting in inevitable, sometime severe, image quality degradation.

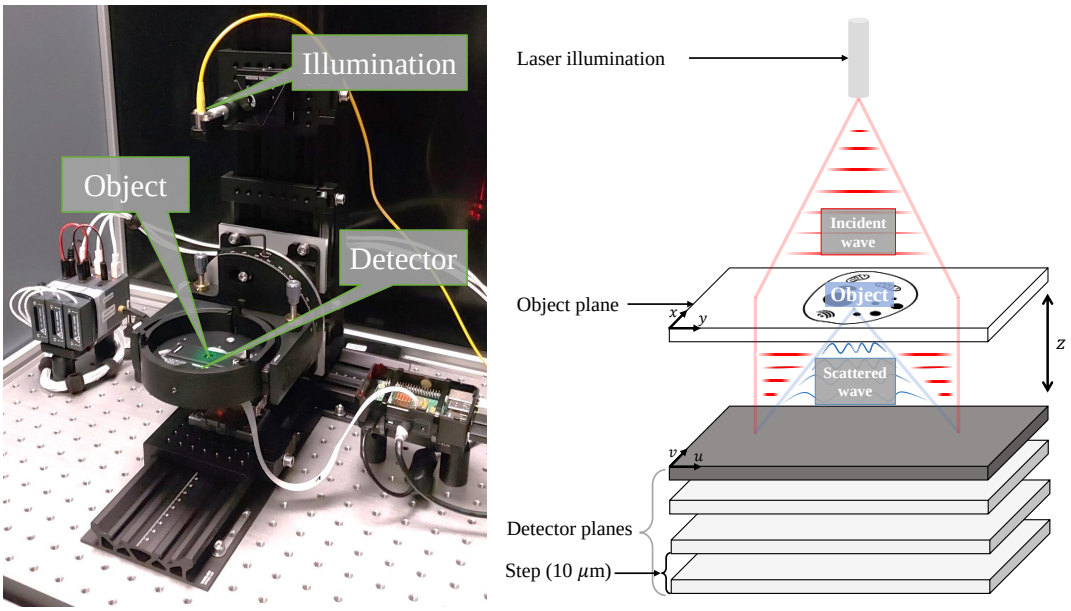


FIGURE 3.1: Lens-free in-line holographic setup: (Left) Digital In-line Holographic Microscope (DIHM) used in this work. (Right) schematics of the different DIHM components.

Recent advancements have demonstrated the efficacy of learning-based methods in tackling holographic reconstruction, both in supervised [Riv+18; Che+22; Che+23a] and unsupervised [Hua+23] manners and to suppress, to some degree, the *twin image* artifacts. The scarcity of real large-scale holographic measurements, coupled with ground truth complex transmission data, has led most approaches in the literature to rely primarily on synthetic data. However, due to the significant domain gap between synthetic and real holographic measurements, achieving model transferability to different domains remains challenging and no approach with effective generalization capabilities have been demonstrated so far.

This work address for the first time this issue and proposes a comprehensive and versatile framework based on an unrolled deep learning architecture inspired by a model-based reconstruction strategy that is robust to domain changes. This is made possible through the proposed model's interpretability: in this way one can leverage large scale synthetic data to effectively learn the inverse holographic image formation model and showcase outstanding generalization ability to the real world domain without any explicit adaptation. Furthermore, a joint framework for in-line holographic image reconstruction and spatial image super-resolution is introduced, where complimentary spatial information in the form of aliasing introduced by sub-pixel displacements between consecutive holographic measurements is explicitly exploited for the benefit of image super-resolution. The main contributions of this work are therefore the following:

- This work introduces a versatile framework that seamlessly integrates in-line holographic image reconstruction with spatial super-resolution and supports an extensive refocusing range.

- Generative models are used to generate large scale synthetic data leveraged to learn the inverse model underlying in-line holographic image formation within an interpretable manner which ensures robust generalization capabilities well beyond the data distribution of the training domain.
- The proposed approach is validated on standard datasets and also through real world samples imaged with a costum made Digital in-line Holographic Microscope (DIHM): it achieves high-quality reconstruction and demonstrates the framework's efficacy and practical utility in real-world applications.

3.2 Prior Art

3.2.1 Iterative Phase Retrieval Algorithms

These algorithms are typically based on error reduction and commonly used in digital in-line holography to reconstruct the latent complex field from real-valued holographic measurements. Gerchberg and Saxton [GS94] first proposed an alternating field projections approach between the object and detector planes while enforcing support constraints, such as positive absorption profiles within defined areas on the object plane, and intensity constraints on the detector plane, where the modulus square of the field must match the measured intensity (see figure 3.2). With sufficient number of iterations the field typically converges towards the latent one. Fienup [Fie78] later proposed some modifications to the original method of [GS94] resulting in the Hybrid Input Output (HIO) variant which incorporates a feedback parameter to relax the hard support constraint on the object plane, resulting in significantly better reconstruction quality with faster convergence. Indeed, it can be shown that such error reduction approaches are special cases of an inverse problem solving framework [Mom+19] which can be prone to bad local minima and in some cases divergence. Fienup's modification helps avoid possible stagnation by further regularizing the possible solution space, HIO has been widely used to this day in the literature and it serves as a baseline method, with variants using single or multiple input holograms captured at different heights referred to as "SH-PR" or "MH-PR", respectively. Iterative methods leveraging prior knowledge of the target sample, such as sparsity [Den+09b; Zha+18], have demonstrated the ability to produce cleaner images by enforcing such constraints. Moreover, iterative solvers help mitigate the undesirable artifacts of the *twin image*, as demonstrated in previous studies. For instance, Zhang et al. [Zha+18] devised a compressive sensing approach employing an iterative shrinkage/thresholding strategy to gradually reduce such artifacts. Latychevskaia et al. [LF07] achieved twin image-free in-line holography through an iterative alternating projections approach without enforcing prior knowledge on the object of interest. Chen et al. [CWH22] proposed an iterative method for holographic image reconstruction and computational refocusing optimizing a least squares problem with plain gradient descent steps. Niknam et

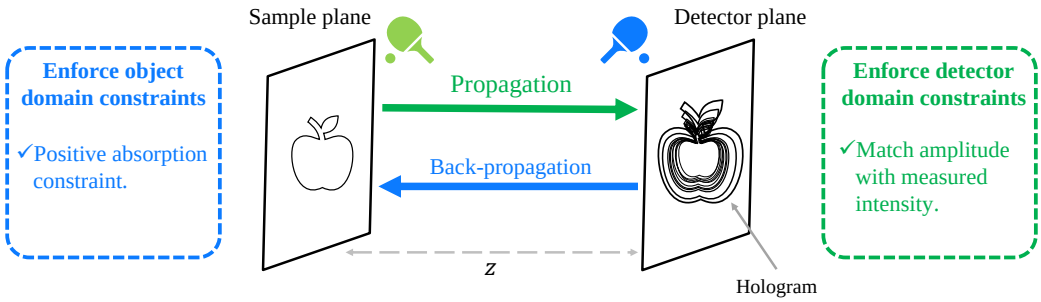


FIGURE 3.2: Alternating projections in between sample and detector planes

al. [NQL21] and Chen et al. [Xiw+24] employed an untrained neural network with randomly generated weights as a natural image prior within a model-based reconstruction framework. Adversarial iterative techniques have also been investigated by Chen et al. [Che+23b] where the authors used a generative network to learn the inverse image formation model of in-line holography, a discriminator is then used to distinguish between the original hologram and the one simulated using the predicted complex field.

3.2.2 Learning-based methods

These methods offer inherent immunity to the *twin image* issue. Networks trained using pairs of holographic images and sharp ground truth data learn a direct mapping between the two sets, circumventing the need to model the underlying physics of holographic image formation. Once trained successfully, these methods can produce twin image-free and sharp reconstructions. However, this straightforward black box approach may encounter unexpected failure cases when presented with new data exhibiting different statistics than that seen during training. Recent work by Chen at al. [Che+22; Che+23a] demonstrated promising model transferability by training a network based on Fourier operators [Li+20] using real data from a specific biological tissue type and testing it on different other types. Despite this advancement, such approaches may struggle to generate high-quality images when confronted with test samples that significantly differ from the training data or originate from entirely different distributions. The work of Chen et al. [Che+23a] is closely related to the proposed approach in this thesis, as it also performs spatial super-resolution. However, while the exact methodology employed by [Che+23a] to perform such task remains undisclosed, several key distinctions set this work apart: (i) image alignment and registration is explicitly incorporated to leverage aliased information for spatial super-resolution. (ii) The proposed network architecture prioritizes interpretability, enhancing robustness to new unseen real-world data. (iii) The proposed network operates independently of changes in illumination wavelength and/or detector/object distance, eliminating the need for retraining when

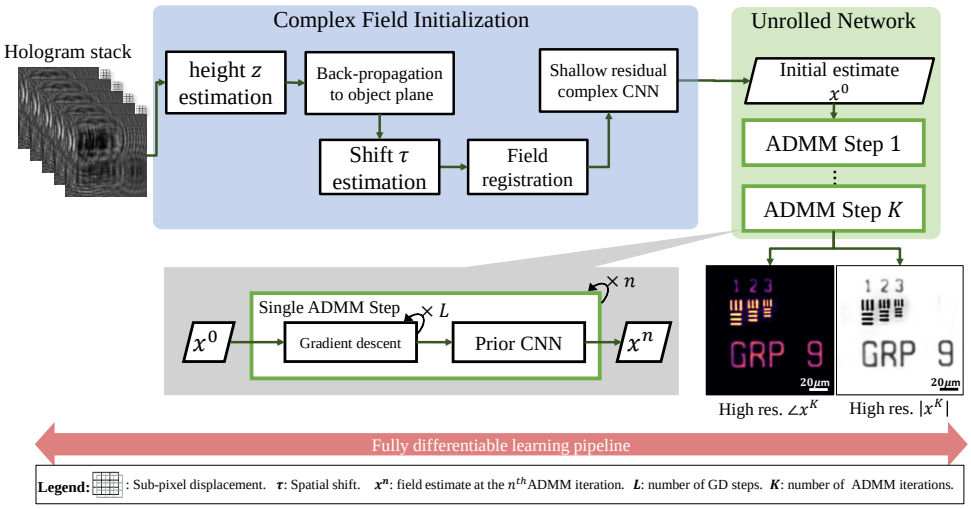


FIGURE 3.3: The overall fully differentiable architecture of HoloADMM: A stack of low resolution noisy holograms captured at different heights used to reconstruct a high-quality and spatially super-resolved complex field.

these variables are altered. Rivenson at al. [Riv+18] introduced a supervised technique where an input hologram is first back-propagated through free space to generate an initial estimate of the latent field which is then fed into a multi-scale CNN that refines it. Huang et al. [Hua+23] proposed a self-supervised learning approach where a network similar to that of Chen et al. [Che+22] is trained using a physics consistency loss with a strategy similar to that of [Che+23b] where the loss is evaluated between the input hologram and a simulated one using the predicted complex field distribution.

3.3 Methodology

In this section the image formation model for in-line holography will be presented along with the problem formulation and the proposed model.

3.3.1 Image Formation Model

The schematics in figure 3.1 illustrate the basic setup of Gabor's in-line lensless holographic imaging system which is use in the DIHM prototype in this work. Given a latent complex transmission field at the object plane $\mathbf{x} \in \mathbb{C}^{hw}$ sampled at high resolution with spatial dimensions $h \times w$, for each height z_i, $i = 1, ..., N$ a hologram can be simulated using the following equation:

$$f_{s,\tau_i,z_i}(\mathbf{x}) = D_s W_{\tau_i} |P_{z_i}\mathbf{x}|^2 \tag{3.1}$$

Where $f_{s,\tau,\mathbf{z}} : \mathbb{C}^{hw} \mapsto \mathbb{R}^{h'w'}$ is the forward in-line holographic image formation model. The latent field $\mathbf{x}$ is propagated to the detector plane using the complex near-field Fresnel propagation kernel P_{z_i} [Foi+08]. The real valued hologram is obtained by calculating the modulus square of $P_{z_i}\mathbf{x}$ which is then spatially warped using the matrix W_τ simulating spatial shifts in the (u, v) plane (see figure 3.1) with sub-pixel accuracy, defined by the set $\tau = \{(\tau_1^u, \tau_1^v), ..., (\tau_N^u, \tau_N^v)\}$. These spatial shifts are necessary for enabling image super-resolution capability, see Section 3.4.3 for more details. D_s is a down-sampling matrix that reduces the image size by a factor s, resulting in dimensions $h' = h/s$ and $w' = w/s$. Notice that the image formation model as described here is non-linear, more general, and physically accurate where the object of interest is assumed to have both absorption and phase shift properties—though absorption can sometimes be negligible for thin and transparent micro-organisms. Additionally, sensor read and shot noise sources are also simulated using the noise model from [Foi+08]:

$$\sigma(k) = \sqrt{\alpha \cdot y(k) + \gamma} \tag{3.2}$$

Where σ is the pixel-dependent standard deviation of the noise level at pixel k, with α and γ representing the variances of shot and read noises, respectively, and $y(k)$ is the clean input pixel value.

3.3.2 Problem Formulation

The aim of this work is to reconstruct a spatially super-resolved complex transmission distribution from a given stack of noisy low-resolution input holograms $H = [\mathbf{h_1}, ..., \mathbf{h_N}]$ captured at N different heights $\mathbf{z} = [z_1, ..., z_N]$ with spatial shifts $\tau = \{(\tau_1^u, \tau_1^v), ..., (\tau_N^u, \tau_N^v)\}$. This problem can be reformulated as a regularized least squares minimization:

$$(\tilde{\mathbf{x}}, \tilde{\tau}) = \arg\min_{\mathbf{x},\tau} \frac{1}{2} \|H - f_{s,\tau,\mathbf{z}}(\mathbf{x})\|_2^2 + \beta\Psi(\mathbf{x}) \tag{3.3}$$

Where Ψ is a regularizer that constrains the possible solution space on the distribution of the latent field $\mathbf{x}$ and β is a discrepancy parameter controlling the strength of such regularization. Accurate prediction of $\mathbf{z}$ is important in order to reconstruct $\mathbf{x}$ as the field at the detector plane can be propagated to the object plane and vice-versa using $P_\mathbf{z}$ or its conjugate $P_\mathbf{z}^*$. The value of $\mathbf{z}$ can be accurately estimated using any computational refocusing technique from the literature [Tam+17]. Spatial shifts τ need to be estimated with sub-pixel accuracy for multi-frame image registration, thereby enabling spatial image super-resolution. The optimization framework as expressed in Eq. 3.3 is non-linear in all optimization variables $(\tilde{\mathbf{x}}, \tilde{\tau})$ and can be solved by iteratively optimizing for one variable at a time while keeping the other one fixed,

i.e., by alternating the following (A) and (B) steps:

$$(A)\ \ \tilde{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2} \|H - f_{s,\tau,\mathbf{z}}(\mathbf{x})\|_2^2 + \beta\Psi(\mathbf{x}) \tag{3.4}$$

$$(B)\ \ \tilde{\tau} = \arg\min_{\tau} \frac{1}{2} \|H - f_{s,\tau,\mathbf{z}}(\mathbf{x})\|_2^2 \tag{3.5}$$

3.3.3 HoloADMM: End-to-end Learning For QPI

Solving for the latent field (A): Eq. 3.4 is a regularized non-linear least squares with no closed-form solution for $\tilde{\mathbf{x}}$: a good approximation can be obtained iteratively with a proper image prior Ψ such as Total Variation [ROF92] or other natural image priors. Choosing a proper Ψ to solve Eq. 3.4 is not trivial: depending on the target scene properties, multiple possible image priors can be used. In this work, a deep convolutional neural network is used to learn a suitable prior agnostic to individual scene properties owing to the large representation capacity of deep networks. To this end, a variable splitting technique is used to solve Eq. 3.4 namely the Alternating Direction Method of Multipliers (ADMM)[GM76] where an auxiliary variable $\mathbf{v}$ is introduced such that:

$$(\tilde{\mathbf{x}}, \tilde{\mathbf{v}}) = \arg\min_{\mathbf{x},\mathbf{v}} \frac{1}{2} \|H - f_{s,\tau,\mathbf{z}}(\mathbf{x})\|_2^2 + \beta\Psi(\mathbf{v})\ \ s.t\ \ \mathbf{x} - \mathbf{v} = 0 \tag{3.6}$$

Note that Eq. 3.6 is now a constrained version of the previous formulation in Eq. 3.4 where the data-fidelity and prior terms are no longer coupled and consequently they can be solved for separately, Eq. 3.6 can be further split into multiple sub-problems by first retrieving its scaled augmented Lagrangian:

$$\mathcal{L}_\rho^{\text{ADMM}}(\mathbf{x}, \mathbf{v}, \mathbf{u}) = \frac{1}{2} \|H - f_{s,\tau,\mathbf{z}}(\mathbf{x})\|_2^2 + \beta\Psi(\mathbf{v}) + \frac{\rho}{2} \|\mathbf{x} + \mathbf{u} - \mathbf{v}\|_2^2 + \frac{\rho}{2} \|\mathbf{u}\|_2^2 \tag{3.7}$$

where $\mathbf{u}$ is the scaled Lagrange multiplier and ρ is a penalty term for the constraint in Eq. 3.6 forcing the final estimate $\mathbf{v}$ to be as close as possible to the true solution $\mathbf{x}$. To minimize Eq. 3.6 the saddle point of Eq. 3.7 needs to be found by iteratively solving the following three sub-problems:

$$\mathbf{x} \leftarrow \arg\min_{\mathbf{x}} \mathcal{L}_\rho^{\text{ADMM}}(\mathbf{x}, \mathbf{v}, \mathbf{u}) = \arg\min_{\mathbf{x}} \frac{1}{2} \|H - f_{s,\tau,\mathbf{z}}(\mathbf{x})\|_2^2 + \frac{\rho}{2} \|\mathbf{x} + \mathbf{u} - \mathbf{v}\|_2^2 \tag{3.8}$$

$$\mathbf{v} \leftarrow \arg\min_{\mathbf{v}} \mathcal{L}_\rho^{\text{ADMM}}(\mathbf{x}, \mathbf{v}, \mathbf{u}) = \arg\min_{\mathbf{v}} \frac{\rho}{2} \|\mathbf{x} + \mathbf{u} - \mathbf{v}\|_2^2 + \beta\Psi(\mathbf{v}) \tag{3.9}$$

$$\mathbf{u} \leftarrow \mathbf{u} + \mathbf{x} - \mathbf{v} \tag{3.10}$$

The full model architecture is shown in figure 3.3: in each ADMM iteration (shown in the grey box), $\mathbf{x}$ in Eq. 3.8 is updated using multiple steps of a plain gradient descent or any other gradient based update rule, e.g., conjugate gradient. Since $\mathbf{x} = \mathbf{a} + i\mathbf{b} \in \mathbb{C}$, gradients are calculated using Wirtinger derivatives [MT95]

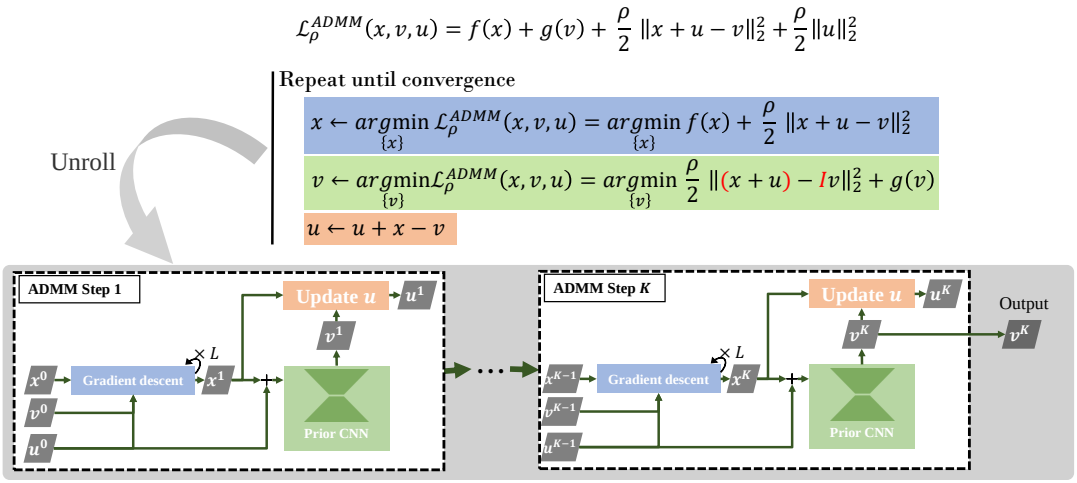


FIGURE 3.4: Detailed implementation of each unrolled ADMM iteration.

where $\mathbf{x}$ is updated via:

$$\mathbf{x} \leftarrow \mathbf{x} - \alpha \cdot \frac{\partial}{\partial \mathbf{x}^*} \mathcal{L}_\rho^{\text{ADMM}}(\mathbf{x}, \mathbf{v}, \mathbf{u}) \tag{3.11}$$

Where $\frac{\partial}{\partial \mathbf{x}^*} = \frac{1}{2}(\frac{\partial}{\partial \mathbf{a}} + i\frac{\partial}{\partial \mathbf{b}})$, $\mathbf{x}^*$ is the complex conjugate of $\mathbf{x}$, and α is a learning rate. The gradient calculation requires defining the backward model of $f_{s,\tau,\mathbf{z}}$ which is denoted in this work with $f^b_{s,\tau,\mathbf{z}}$: It can be approximately implemented by first upsampling all N low resolution input holograms, back-propagating each hologram in the stack using $P^*_{z_i}$, $i = 1,...,N$ back to the object/sample plane, and aligning the resulting complex images. Finally, the final back-propagated output $\mathbf{x}^b$ is obtained by averaging all N aligned complex fields. Eq. 3.9 can be viewed as a simple denoising problem with the identity matrix I as the forward model and a noisy input $x+u$, where the target is to estimate a clean complex field v. In principle, any plug-and-play denoiser should be suitable to solve Eq. 3.9. However, to achieve better performance and enable end-to-end learning, a trainable convolutional neural network is used as a denoiser that acts as a learned image prior which in this case is a ResUnet architecture [ZLW18]. Finally, the update step of $\mathbf{u}$ is straightforward.

figure 3.4 shows a detailed implementation of the update rule for any given intermediate estimate $\mathbf{x}$ where multiple steps of gradient descent are first performed to solve Eq. 3.8 followed by a denoising step to solve Eq. 3.9. Notice that the weights of the prior ResUnet architecture are shared among all ADMM iterations which are unrolled to form the overall network architecture of the proposed HoloADMM. The weights are learned in a supervised end-to-end manner together with all the other hyper-parameters namely the scaled Lagrange multiplier ρ from Eq. 3.7 and the gradient descent learning rate α from Eq. 3.11.

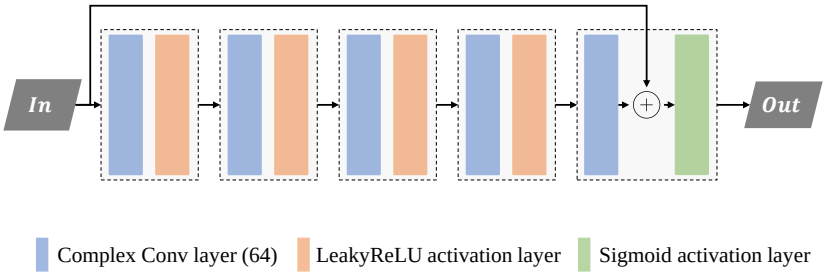


FIGURE 3.5: Shallow residual network.

Solving for spatial shifts (B): In principle any image registration algorithm can be used to estimate spatial shifts $\tau = \{(\tau_1^u, \tau_1^v), ..., (\tau_N^u, \tau_N^v)\}$ between a reference frame (hologram closest to the object plane) and all the other frames. Note that the registration is not performed on the raw input holograms because diffraction patterns are different due to field propagation since the distance between the object and detector planes changes with each capture. The registration is instead carried out on the images of the back-propagated stack of N holograms. In this work, a fast FFT based alignment approach with arbitrary sub-pixel accuracy [GSTF08] is used to align the images in the input stack.

Complex shallow network: Since the target domain is complex by nature, the initial estimate ($\mathbf{x}^0 \in \mathbb{C}$) undergoes further processing through a shallow complex convolution network [Gub16] in order to learn close correlations between its real and imaginary components without the need to separate them into two distinct channels. Such network, depicted in figure 3.5, is designed with a residual connection linking the input and output distributions, enhancing information flow and reducing noise in $\mathbf{x}^0$: it has multiple convolution layers with complex kernels in $\mathbb{C}$ thus the learned weights are complex in nature. Complex convolution neural networks [Gub16] are suitable to learn correlations in the complex domain where the inputs as well as the learned weights are in $\mathbb{C}$. Let $\mathbf{x} \in \mathbb{C}$ be a complex transmission field expressed as $\mathbf{a} + i\mathbf{b}$, and $\omega \in \mathbb{C}$ a complex learnable convolution kernel expressed as $\omega_R + i\omega_I$. A complex convolution operation can therefore be seen as a combination of four different real ones:

$$\begin{aligned} \omega * \mathbf{x} &= (\omega_R + i\omega_I) * (\mathbf{a} + i\mathbf{b}) \\ &= (\omega_R * \mathbf{a} - \omega_I * \mathbf{b}) + i(\omega_R * \mathbf{b} + \omega_I * \mathbf{a}) \end{aligned} \tag{3.12}$$

The shallow residual network used in this work is depicted in figure 3.5. It has 5 convolutional layers, each with 64 learnable filters followed by a LeakyRELU activation layer except for the last one that is followed by a Sigmoid activation layer. The residual connection prompts the network to perform image enhancement, as shown in figure 3.6, where the output is just a sharper and enhanced version of the

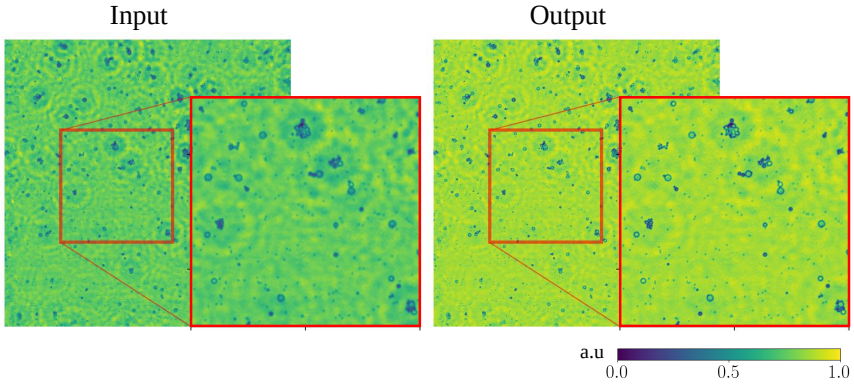


FIGURE 3.6: Input and output images (from a real beads hologram) of the shallow complex CNN. The network learned to enhance image quality by producing sharper details.

input image. Recall that the output of such network serves as an initial guess to the unrolled network based on the ADMM solver that is the next step.

Initialization: As depicted in the initialization block in figure 3.3 (blue box), after estimating the relative spatial displacements using the back-propagated fields, $\mathbf{x}^0$ is obtained by applying the backward model $f^b_{s,\tau,\mathbf{z}}$ as described before and feeding the resulting complex field to the shallow residual complex network. $\mathbf{u}^0$ and $\mathbf{v}^0$ are set to 0. The initial value for the learning rate α of the plain gradient descent step in Eq. 3.11 is set to 0.01 and ρ in Eq. 3.7 is set to 0.1, recall that both of these variables are learned in an end-to-end fashion. The number of ADMM steps/iterations $n = 5$ which are unrolled, and the number of gradient descent steps in each iteration is $L = 3$.

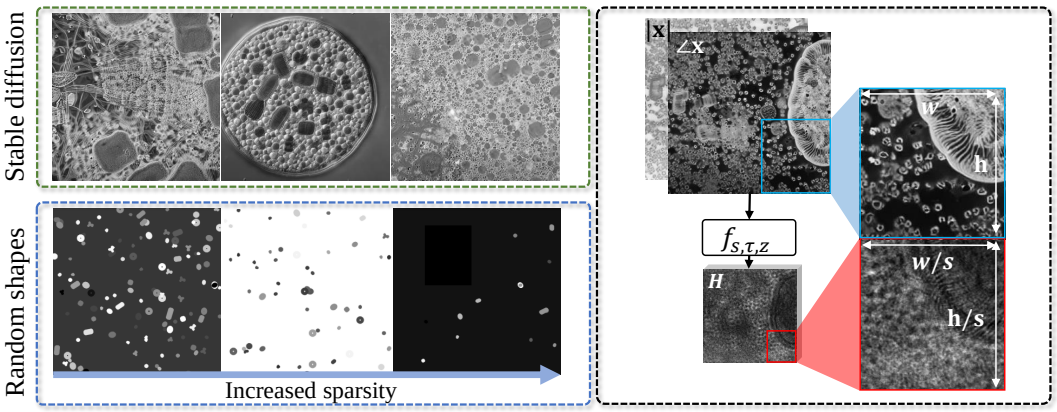


FIGURE 3.7: Synthetic data generated using a stable diffusion model (top left) and a software that generate random shapes with different sparsity levels (bottom left). An input latent field and its simulated hologram (right).

3.3.4 Datasets and Training Details

The lack of real holographic datasets, along with ground truth complex fields, is the main limitation hindering the development of learning-based reconstruction models. Self-supervised approaches, [Hua+23; NQL21] where the loss is evaluated between a real captured hologram and a simulated one using the predicted phase and amplitude distributions of the latent complex field, suffer from three major drawbacks: (i) The problem formulation is severely ill-posed since multiple possible solutions might correspond to the same measurement, eventually leading to inaccurate reconstruction that does not necessarily correspond to the true target, (ii) designing effective self-supervised loss functions can be challenging. These loss functions need to capture relevant characteristics of the data and the reconstruction task, which may not always be straightforward to define, (iii) The real forward model, if not carefully designed, will further contribute to low reconstruction quality even with small simulation inaccuracies. Large scale real data with ground truth complex distributions is not trivial to collect, in fact, ground truth data used in the literature is obtained by the MH-PR algorithm [Fie78] or other closely related variants which imposes an upper-bound hindering the true reconstruction capability of any network. One can therefore argue that learning an accurate inverse model through large-scale pixel-accurate synthetic data is crucial for high-quality reconstruction. To this end, a large number of synthetic microscopic data is generated featuring images with varying complexity and sparsity. A generative stable diffusion model [Rom+22] fine-tuned on microscopic images[1] is used to generate dense interconnected samples with fine spatial details, in addition, samples with a varying degree of sparsity were obtained using a simple software that generates a random number of simple shapes in a canvas. figure 3.7 shows some samples generated using the two modalities; sparse as well as dense data samples are generated to account for the real nature of microscopic images where individual or few cells as well as dense connective tissues might be present in a given image. Phase and amplitude distributions are obtained from a single gray-scale image $\mathbf{I}$ by $A = |\mathbf{x}| = e^{\omega \times \mathbf{I}}$, $\Phi = \angle \mathbf{x} = \pi \mathbf{I}$, where $\omega \leq 0$ is a weight determining the degree of transparency of the sample, no amplitude information (or fully transparent image) corresponds to $\omega = 0$, data is simulated with $\omega = -1.6$ to favor highly transparent samples. The simulated dataset contains more than 100k different training samples. As shown in figure 3.7 (right) input low resolution noisy holograms are coupled with high-resolution clean phase and amplitude images. During training, wavelengths are chosen randomly from [440 nm, 530 nm, 638 nm] and used to simulate each hologram along with a broad refocusing range from 0.5 mm up to 1.0 mm with a step size of 10μm. This dataset is challenging because a single object can have many corresponding holographic measurements each with different illumination wavelength and/or refocusing distance, forcing any learning-based approach to effectively learn the inverse model which is agnostic to

[1] https://huggingface.co/Fictiverse/Stable_Diffusion_Microscopic_model

TABLE 3.1: Quantitative comparison results on D_i and D_o synthetic test data without spatial super-resolution. (†) is a self-supervised approach.

Method	RMSE↓ ($\times 10^{-3}$)				PSNR↑				SSIM↑			
	A		Φ		A		Φ		A		Φ	
	D_i	D_o	D_i	D_o	D_i	D_o	D_i	D_o	D_i	D_o	D_i	D_o
Wu et al. [Wu+21]	-	-	20.31	12.47	-	-	17.160	19.163	-	-	0.364	0.640
Huang et al.† [Hua+23]	-	-	7.81	12.50	-	-	14.271	17.715	-	-	0.602	0.549
Ren et al. $\times 1$ [RXL19]	-	-	8.11	4.31	-	-	21.343	23.822	-	-	0.647	0.700
Rivenson et al. [Riv+18]	5.34	6.77	9.43	9.83	23.113	23.418	20.628	21.987	0.736	0.766	0.681	0.703
Chen el at. [Che+22]	2.23	9.70	4.33	15.29	27.318	24.397	24.451	21.402	0.853	0.782	0.820	0.672
HoloADMM $\times 1$	**1.23**	**0.96**	**2.42**	**1.55**	**29.373**	**30.198**	**26.410**	**28.092**	**0.923**	**0.905**	**0.902**	**0.877**

changes in the input distribution. Sub-pixel shifts are randomly simulated in the range of ± 3 pixels. HoloADMM is trained using the Mean Squared Error (MSE) as loss for 100 epochs with the Adam optimizer and a learning rate of $1e^{-4}$.

3.4 Results and Discussions

Results on synthetic as well as real in-line holographic data are presented in this section along with quantitative and qualitative comparisons with other competing approaches.

Literature reproduction The work of Wu et al. [Wu+21] has been reproduced using the Tensorflow implementation provided in `https://github.com/THUHoloLab/Dense-U-net`. Huang et al. [Hua+23] phase-only "GedankenNet" has been reproduced using the author's official implementation provided in `https://github.com/PORPHURA/GedankenNet` and subsequently the work of Chen el at. [Che+22] has been reproduced from the same code base. The network architecture used in Ren et al. [RXL19] has been reproduced from the implementation details provided in their paper. Similarly, the work of Rivenson et al. [Riv+18] has been reproduced using the network implementation details provided in the paper and the supplementary materials. The work of Niknam et al. [NQL21] was reproduced from the official implementation provided in `https://github.com/farhadnkm/DCOD`. The authors of Chen et al. [Che+23b] and Chen et al. [Xiw+24] kindly provided code implementation of their two approaches used in this work.

3.4.1 Synthetic Holographic Data

HoloADMM and its competitors are trained exclusively on synthetic data, as detailed in Section 3.3.4, and evaluated on both inner and outer synthetic test sets denoted as D_i and D_o respectively: the former comprises synthetic images generated using the approach outlined in Section 3.3.4, while the latter contains a handful of classic test targets taken from the Set14 dataset [HSA15] (the chosen image indices are $[1, 2, 9, 10, 12]$). HoloADMM is trained using $N = 10$ holograms, yet it is able to infer complex fields from an arbitrary number of input holograms, provided enough memory. As demonstrated below, the proposed model consistently exhibits superior

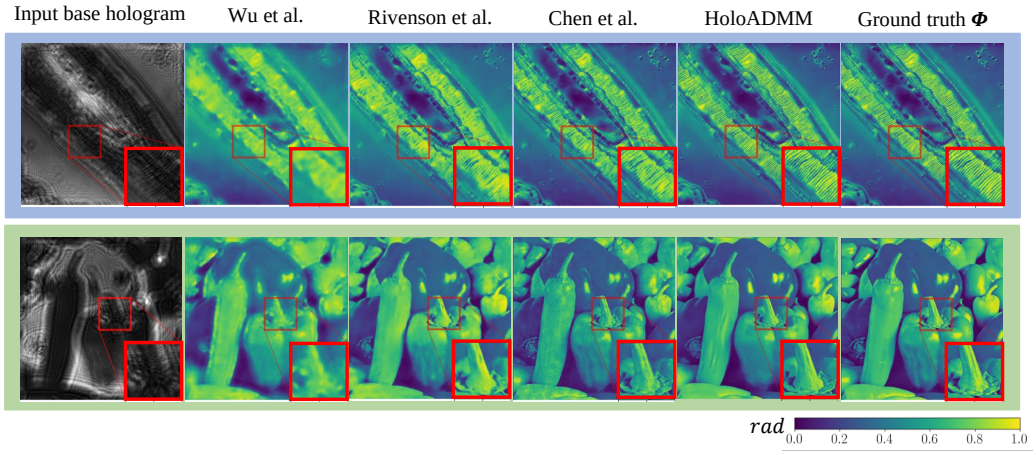


FIGURE 3.8: Reconstructed Φ from some selected synthetic holographic samples taken from D_i and D_o.

TABLE 3.2: Quantitative comparison results on on D_i and D_o synthetic test data with spatial super-resolution.

Method	SR factor	RMSE↓ ($\times 10^{-3}$)				PSNR↑				SSIM↑			
		A		Φ		A		Φ		A		Φ	
		D_i	D_o	D_i	D_o	D_i	D_o	D_i	D_o	D_i	D_o	D_i	D_o
Ren et al. [RXL19]	×4	-	-	23.76	75.49	-	-	16.531	14.159	-	-	0.284	0.520
HoloADMM	×4	**2.46**	**1.32**	**4.59**	**1.94**	**26.544**	**29.313**	**23.764**	**27.626**	**0.806**	**0.827**	**0.769**	**0.791**
HoloADMM	×8	**7.13**	**8.53**	**12.35**	**14.94**	**21.910**	**22.938**	**19.442**	**20.794**	**0.554**	**0.663**	**0.488**	**0.598**

reconstruction quality compared to the state-of-the-art, even when provided with just a single hologram as input or when the input is heavily down-sampled. Quantitative metrics reported in table 3.1 demonstrate HoloADMM's efficacy on both inner and outer datasets, outperforming competing approaches by considerable margins according to all reported metrics with an average PSNR improvement of over 4dB on phase (Φ) images compared to the second best approach on D_o. Performances are also confirmed by figure 3.8 where visual inspection reveals HoloADMM's capability to preserve image details while effectively suppressing sensor noise, resulting in clean and sharp phase images in contrast to competing methods, note that [Che+22] fails to suppress signal-dependent noise and [Wu+21; Riv+18] suffer from blur artifacts. When considering also joint spatial super-resolution, HoloADMM is trained using a decimated stack of holograms by a factor of ×4 with an input shape of 128×128 pixels, yet it achieves superior phase image quality, as depicted in figure 3.10, compared to other competitors trained using high-resolution inputs (512×512 pixels). Table 3.2 quantitatively corroborates these results: HoloADMM reaches an SSIM value of 0.791 with ×4 down-sampled inputs, compared to 0.703 obtained by [Riv+18] with full-resolution input holograms.

Additional qualitative comparison results on synthetic data can be seen in figure 3.9 where some sample reconstructions from HoloADMM as well as other approaches from the literature are shown on the inner test set with clear advantage in

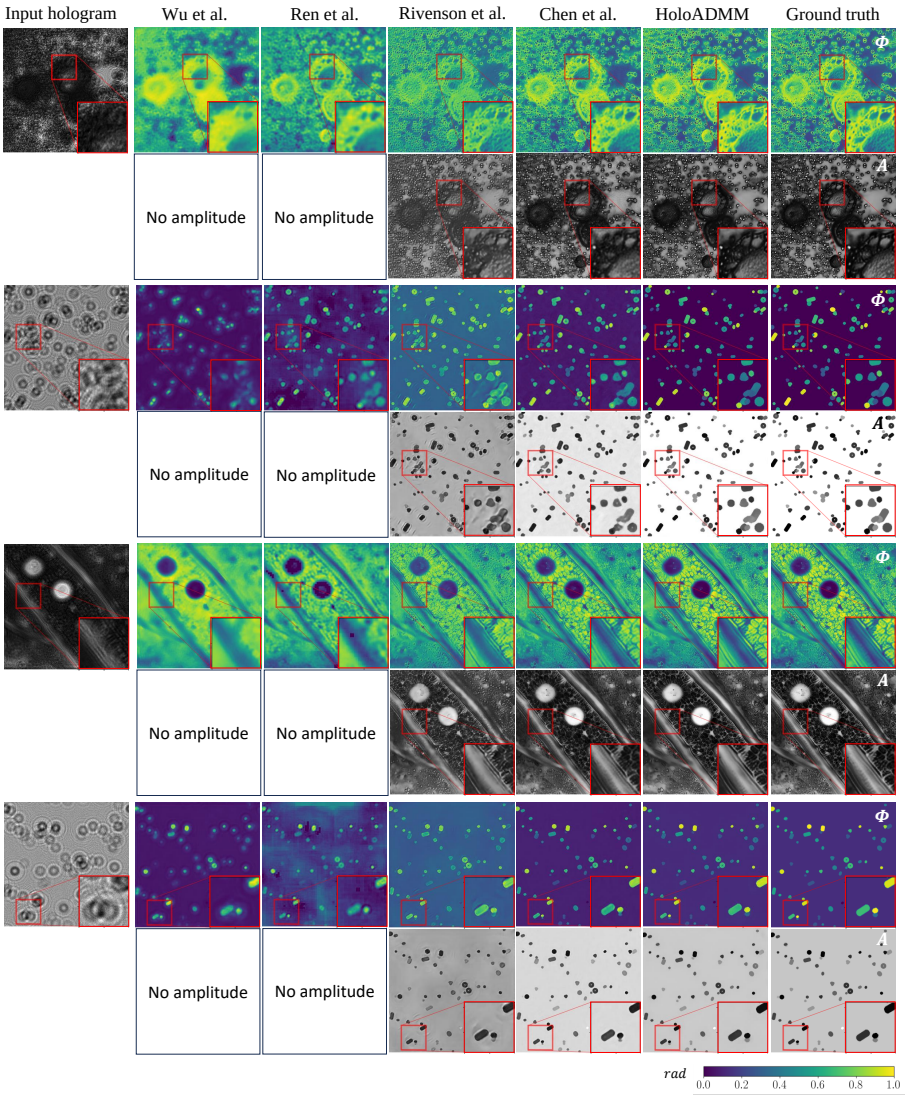


FIGURE 3.9: Reconstructed amplitude and phase images from synthetic holographic data without spatial super-resolution.

terms of visual image quality of both predicted amplitude and phase distributions. figure 3.11 shows some reconstruction results with an extreme super-resolution factor of $\times 8$ on synthetic holographic data data with an input shape of 64×64 pixels and a target resolution of 512×512 pixels. Note that even with 64 fold decimated data, HoloADMM is able to restore most of the scenes' details in this challenging scenario.

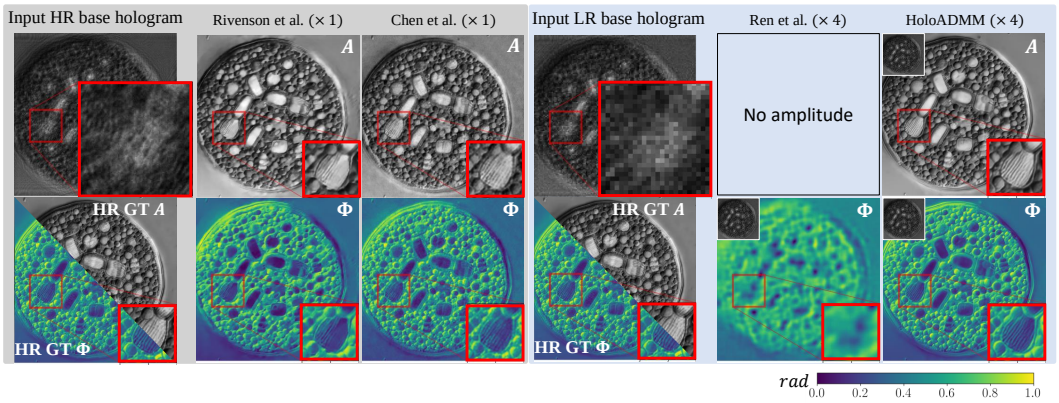


FIGURE 3.10: Reconstructions with ×4 SR for HoloADMM and [RXL19] and ×1 for [Che+22; Riv+18].

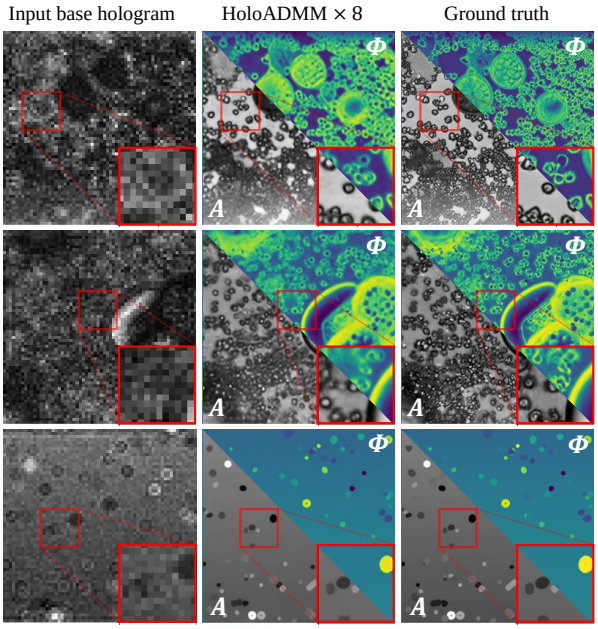


FIGURE 3.11: Reconstructed amplitude and phase images with a spatial super-resolution factor ×8.

Ren et al. [RXL19] also performs spatial super-resolution via sub-pixel convolutions [Shi+16]. Quantitative as well as qualitative results in table 3.2 and figure 3.10 demonstrate that the proposed model produces sharper images, preserving high-frequency details and outperforming [RXL19] for ×4 super-resolution. Even with an extreme factor of ×8, HoloADMM is capable of yielding reasonable quantitative results, reported in table 3.2, in contrast to [RXL19], which fails to generate meaningful image data.

TABLE 3.3: Quantitative results (PSNR) on some standard test targets. The table compares the PSNR achieved by the proposed approach with other iterative methods. Computation times refer to an NVIDIA A6000 except for (*) that uses the CPU only.

Method	# holograms	Time (s)	`Airplane`		`Barbara`		`Baboon`	
			A	Φ	A	Φ	A	Φ
Niknam et al. [NQL21]	1	529	21.019	11.020	16.127	14.261	17.425	13.583
Chen et al. [Che+23b]	1	990	18.297	17.4709	14.720	12.051	15.403	14.960
Chen et al. [Xiw+24]	1	797	17.188	18.012	20.047	17.179	15.504	14.744
SH-PR [Fie78]	1	60*	16.307	9.085	17.540	12.937	14.934	12.753
MH-PR [Fie78]	10	100*	19.465	11.295	19.900	19.565	20.596	18.567
HoloADMM ×1	1	**0.29**	29.178	24.828	25.261	23.959	23.800	21.454
HoloADMM ×1	10	0.58	**32.223**	**28.205**	**28.289**	**26.757**	**27.198**	**25.020**

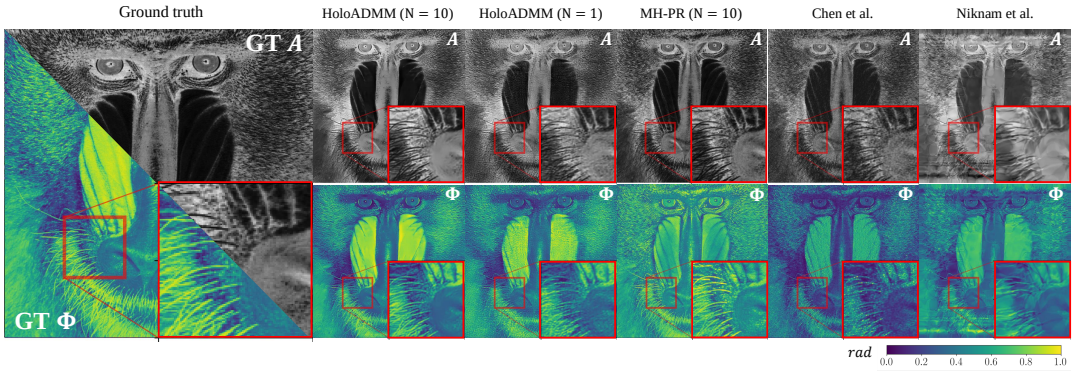
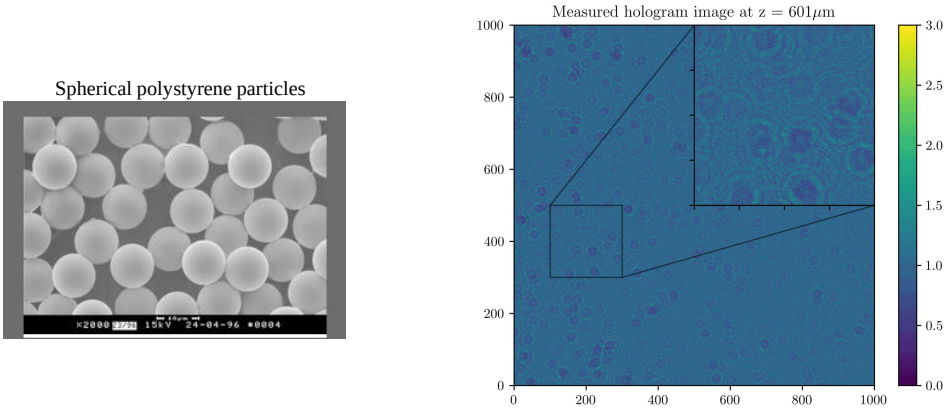
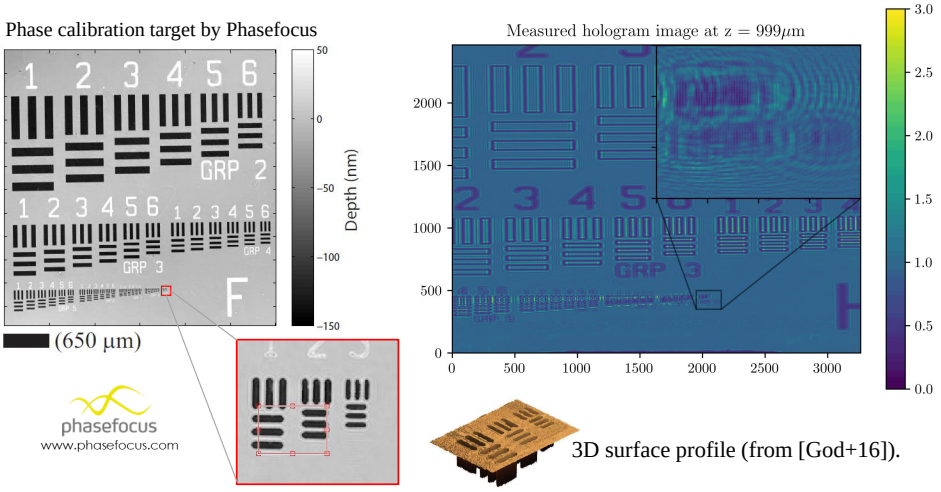


FIGURE 3.12: Reconstruction results of A and Φ on `Baboon` test target.

Additionally, results reported in table 3.3 and figure 3.12 compare the reconstruction performance of HoloADMM, already trained on synthetic data, with other model-based iterative methods on `Airplane`, `Barbara`, and `Baboon` test targets from D_o. For a fair comparison, results using single as well as 10 holograms as input are reported; in both cases, the proposed approach outperforms iterative solvers and effectively suppresses signal-dependent sensor noise, unlike Chen et al. [Che+23b; Xiw+24], where the proposed algorithms tend to fit the noise model as the number of iterations increases. Furthermore, HoloADMM preserves small spatial details, such as the fine whiskers of the baboon in figure 3.12, while avoiding undesirable artifacts like those produced by [NQL21] in an attempt to suppress sensor noise which results in an over-smoothed image. It is noteworthy that iterative approaches typically require a considerable amount of time to produce reasonable results (up to 15 min. for a 512×512 image in the approach of [Che+23b]), while HoloADMM, once trained, can infer complex field distributions in under a second as reported in table 3.3.



(A)



(B)

FIGURE 3.13: Beads and its hologram at 601 μm (A). Phase calibration target and the captured hologram at 999 μm (B).

3.4.2 Real Holographic Data

DIHM Hardware Setup: The system prototype used in this work is depicted in figure 3.1 (left): a narrow-band, wide-range tunable laser light source (a NKT Photonics SuperK laser) combined with a Laser Line Tunable Filter (LLTF) device is used to emit coherent light in the visible range with a bandwidth of 1-2.5 nm. The illumination is directed from the tip of a single-mode fiber, and a detector-filling light cone is achieved by maintaining an illumination-to-object distance of over 200 mm. The used detector is a CMOS sensor (Sony IMX219) with 1.12 μm pixel size. To ensure near-field conditions, where the Fresnel approximation is valid, the object-detector distance is kept below 1 mm. Multi-height acquisition is performed by moving the object in the axial direction using a piezo motor linear stage (CONEX-SAG-LS48P) with a step size of 10 μm, capturing a set of 10 holograms. Using a

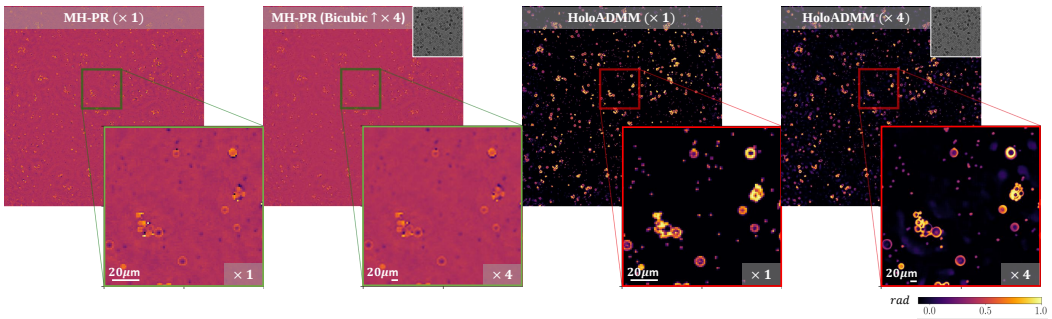


FIGURE 3.14: Reconstructed Φ from real beads holograms with ×1 and ×4 SR. Results from MH-PR are also shown with ×4 resolution using bicubic up-sampling. Input low resolution holograms are shown on the top right corners.

2D piezo motor linear stage, such as the CONEX-SAG-LS48P, it is possible to move the object laterally and anteriorly/posteriorly to capture a series of sub-pixel shifted holograms with a precision of one quarter of a pixel, i.e., 280 nm, enabling spatial super-resolution capability.

Two different test target holograms where captured using the imaging prototype as described above:

- **Polystyrene beads**, a photo of which is shown in figure 3.13a on the left, which mimic real microscopic samples in both structural size and light transmission properties. They have diameters ranging from 0.9 - 9 μm and are equipped with carboxyl functional groups on their surface, allowing them to covalently attach to the cover glass and form a monolayer structure. The cover glass slides are attached to a sticky microchannel slide (IBIDI sticky-Slide I Luer) and the channel is filled with a tailored liquid solution (Immersol). The sample is imaged using an illumination wavelength in the visible range, e.g., 638 nm, to achieve ambiguity-free phase retrieval, prevent phase wrapping, and provide suitable phase contrast, the captured hologram at 601 μm is shown in figure 3.13a on the left.

- **A phase calibration target** "Phasefocus", a photo of which is shown in figure 3.13b on the left, was used to calibrate the DIHM setup. This target is fabricated through Reactive Ion Etching (RIE) of amorphous SiO2, which is patterned via optical lithography, as described in [God+16]. The target includes both phase and amplitude features, which can be imaged using the built DIHM prototype. The phase features are 600 nm deep trenches etched into transparent amorphous SiO2, providing feature sizes that span a wide range of spatial frequencies, from length scales of 2 μm to length scales of 600 μm. The focus of the experiment is on the reconstruction performance of the smallest features, such as GRP 9, which have a structure size of 2 - 10 μm and they are therefore of interest for microscopic application.

All competing models listed in table 3.1 were trained solely on synthetic data tested on the DIHM measurements. As anticipated, almost all models exhibits very

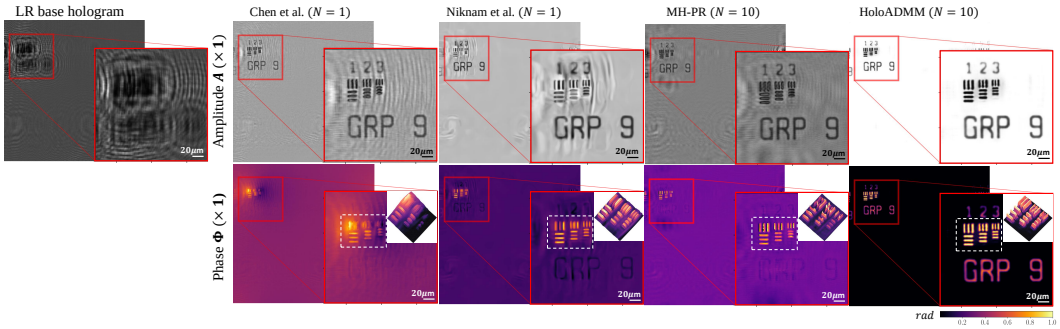


FIGURE 3.15: Reconstructed A and Φ from a real phase calibration target. Highlighted features are invisible in the bright-field domain but exhibit high phase contrast. Reconstructed 3D surfaces from Φ of the etched lines are shown on the top right corner of each zoomed-in region.

poor and sometimes meaningless reconstructions on real data, further highlighting the challenge of model transferability beyond the training/synthetic domain. Qualitative results for those models are shown in figure 3.17. In contrast, HoloADMM is able to produce high-quality reconstructions on different holographic samples for a model trained on synthetic data only and without any form of explicit adaptation. Notably, it not only produces meaningful results but also achieves higher spatial resolution, up to $\times 4$ as shown in figure 3.14 reaching an effective pixel size of 280 nm, alongside results from the standard MH-PR algorithm. Figure 3.15 shows reconstructed GRP 9 features from the phase calibration target, along with outputs from other iterative approaches listed in table 3.3. HoloADMM reconstructs cleaner and sharper phase and amplitude images suppressing noise and undesirable diffraction patterns, revealing fully transparent features with higher contrast and allowing for better quality 3D surface reconstructions from the predicted Φ compared to the other competitors. figure 3.16 contains some sample reconstruction results on real data. It shows the reconstructed complex fields from HoloADMM and competing approaches. Notice that all approaches are trained on synthetic data and tested using the real beads holographic data: while competing methods fail to produce acceptable results and sometimes those results are meaningless, HoloADMM reconstructs significantly better, cleaner, and sharper phase and amplitude images.

figure 3.17 further demonstrates the reconstruction capability of HoloADMM: it depicts three dimensional surface structures obtained from the phase shift estimated from the GPR 9 features of the phase calibration target. The height map h is obtained using the following formula:

$$h = \frac{\lambda \Phi}{2\pi \Delta n} \tag{3.13}$$

Where $\lambda = 638$nm is the wavelength of the illumination laser, Φ is the predicted phase shift, $\Delta n = n_{Cr} - n_{air}$ is the refractive index difference between the phase calibration target material and air. The actual depth of the etched lines is 600 nm. HoloADMM produces well-defined three-dimensional structures, smooth textures and consistent depth values close to the ground truth one compared to other iterative

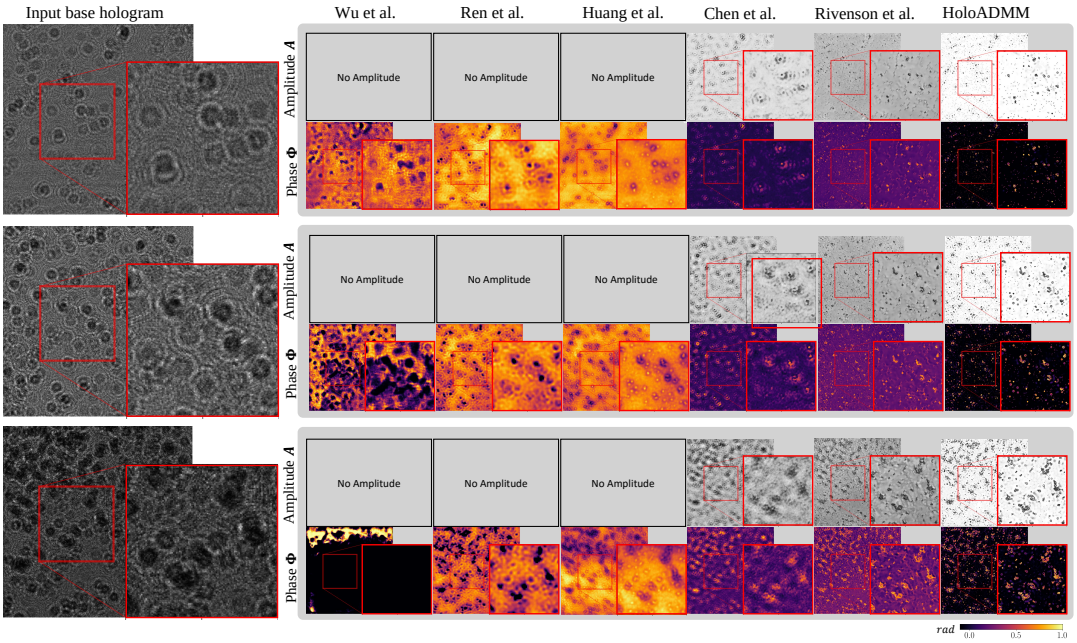


FIGURE 3.16: Reconstructed amplitude and phase images from real beads holograms: All approaches are trained solely on synthetic data. HoloADMM performs $\times 4$ super-resolution (from 512×512 to 2048×2048 pixels)

methods that usually suffer from persistent diffraction artifacts and abrupt structure changes.

3.4.3 Ablation Studies

Different ablation experiments have been conducted on a small subset of the training dataset. The increase in the number of ADMM steps, i.e., the number of unrolled blocks, leads to lower overall loss values, as shown in figure 3.18 (left), but requires more computational resources; in this work the number of iterations is set to $n = 5$. A naive approach that takes a hologram stack as input and predicts the latent field falls short of achieving good performance, even on training data, as shown in figure 3.18 (center), and is unable to generalize beyond that domain. Image registration is crucial for image super-resolution; without such a step, complementary spatial information in the form of aliasing is not exploited, and the spatial quality of the reconstructed images deteriorates, as shown in figure 3.18 (right). Further ablation experiments have been conducted to highlight the model's interpretability, the choice of the prior network, of the loss function, and to assess the robustness to a low number of input holograms and to the change of the refocusing distance.

Learned ADMM hyper-parameters: figure 3.19 shows the evolution of the learned hyper-parameters α and ρ. The learning rate α starts by slightly increasing and then

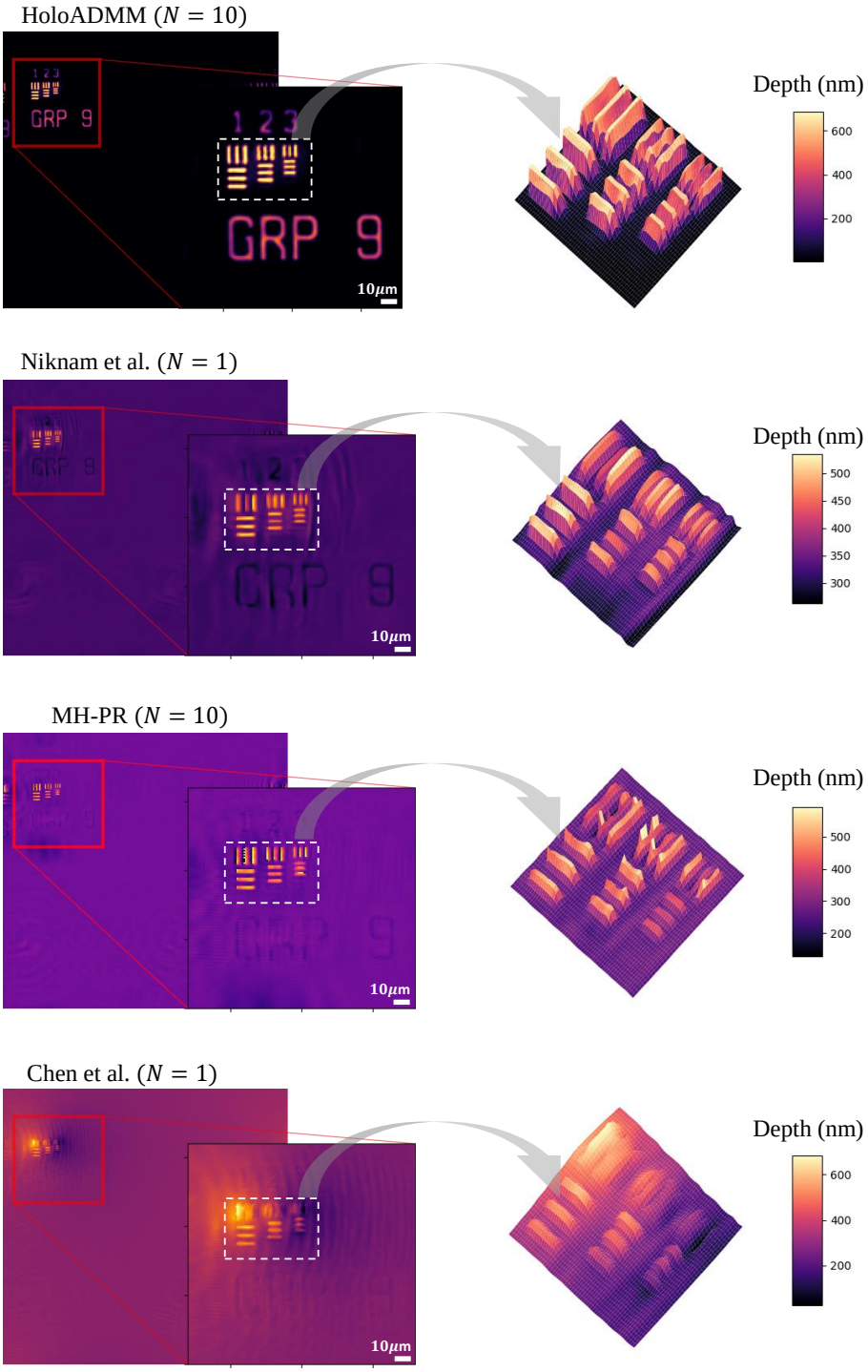


FIGURE 3.17: 3D surface reconstruction of phase calibration target.

FIGURE 3.18: Ablation experiments: (left) with different number of unrolled ADMM steps, (center) with a straightforward approach, (right) without image registration.

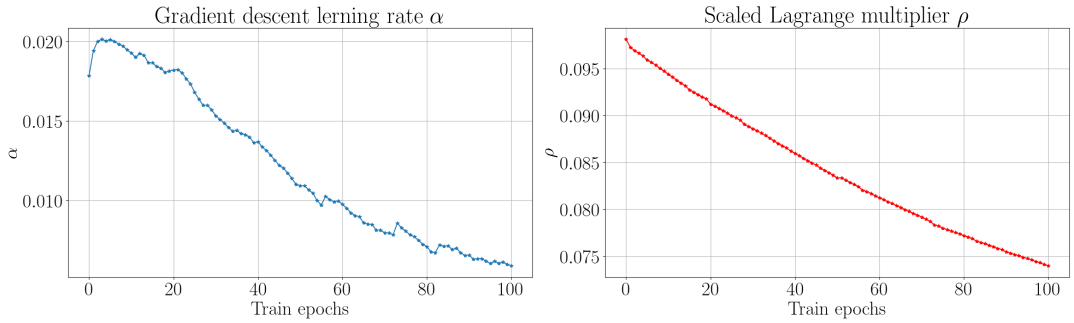


FIGURE 3.19: The evolution of the learning rate α used in Eq. 10 and the scaled Lagrange multiplier ρ in Eq. 6.

decreases as the training progresses. This is expected and makes sense from an optimization point of view: at the start of the training the initial guess is usually suboptimal, thus the gradient descent steps need to have a higher learning rate to reach a lower minima and to possibly avoid stagnation, as the training progresses the network learns to reconstruct better images and the gradient descent step in this case marginally improves upon the initial guess, so the learning rate has to be lower as shown by the curve in figure 3.19 on the left. The Lagrange multiplier ρ is a penalty term and it weights the constraint in Eq. 5 ($\mathbf{v} - \mathbf{x} = 0$). The curve shown in figure 3.19 on the right shows the learned initial value for ρ, note that the network gradually opts for lower starting values for this parameter. Notice also that inside each step, ρ starts from the value in the curve and is then increased following a given schedule [CWE16] as the number of ADMM iterations increases (a linear increase scheme is used by multiplying ρ by 1.1 at each ADMM iteration). In this way, at the end of the optimization iterations the estimate $\mathbf{x}$ would be as close as possible to the auxiliary variable $\mathbf{v}$.

Prior architecture: Several architectures from the literature are tested as the prior part of HoloADMM, namely a UNET [RFB15], a ResUNET [ZLW18], and a ResUNET++ [Jha+19]. Quantitative results are reported in table 3.4. All three architectures achieve competitive results with respect to other approaches from the literature reported previously in table 3.1 given the fact that the input is down-sampled by a factor $\times 4$. This performance indicates that information flow within HoloADMM

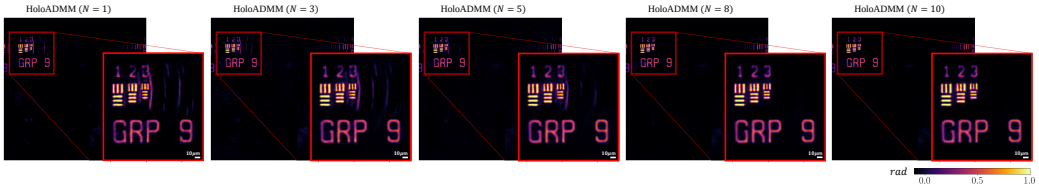


FIGURE 3.20: Reconstruction quality with increased number of input holograms (N).

leads to better reconstruction performance regardless of the prior architecture. Notice that priors with residual connections (ResUNET) and attention modules (ResUNET++) outperform the simple UNET architecture. A larger model size also leads to better quantitative results such as the case of ResUNET (13M) against UNET (8M) and ResUNET++(6M). In this work a ResUNET architecture is used as the neural prior because ResUNET++ is slower due to the attention modules that also impose an upper-bound on the network size given the available GPU memory.

TABLE 3.4: Quantitative results on D_i using different prior architectures with a spatial super-resolution factor $\times 4$.

Prior Network	Size (M)	RMSE↓		PSNR↑		SSIM↑	
		A	Φ	A	Φ	A	Φ
UNET [RFB15]	8	4.33	7.85	24.008	21.350	0.719	0.670
ResUNET++ [Jha+19]	6	3.96	7.24	24.386	21.685	0.741	0.693
ResUNET [ZLW18]	13	**2.46**	**4.59**	**26.544**	**23.764**	**0.806**	**0.769**

Loss functions: HoloADMM is trained using the Mean Absolute Error (MAE) and Mean Squared Error (MSE) loss functions and also a combination of MSE and perceptual loss (VGG loss) [MTZM18]. Quantitative results are reported in table 3.5. The model performs well with the MSE loss function while MAE introduces high frequency artifacts and leads to overall lower quantitative metrics, the combination of MSE and a perceptual loss function does not lead to any noticeable improvement as well.

TABLE 3.5: Quantitative results on D_i using different loss functions with a spatial super-resolution factor $\times 4$.

Loss Function	RMSE↓		PSNR↑		SSIM↑	
	A	Φ	A	Φ	A	Φ
MAE	3.65	6.27	24.756	22.029	0.755	0.711
MSE+VGG	3.55	6.48	24.859	22.177	0.760	0.716
MSE	**2.46**	**4.59**	**26.544**	**23.764**	**0.806**	**0.769**

Robustness: HoloADMM is designed to work best with a stack ($N > 1$) of input holograms which enable multi-frame image super-resolution via alignment and registration. Yet, with a single hologram as input, the model is still able to produce

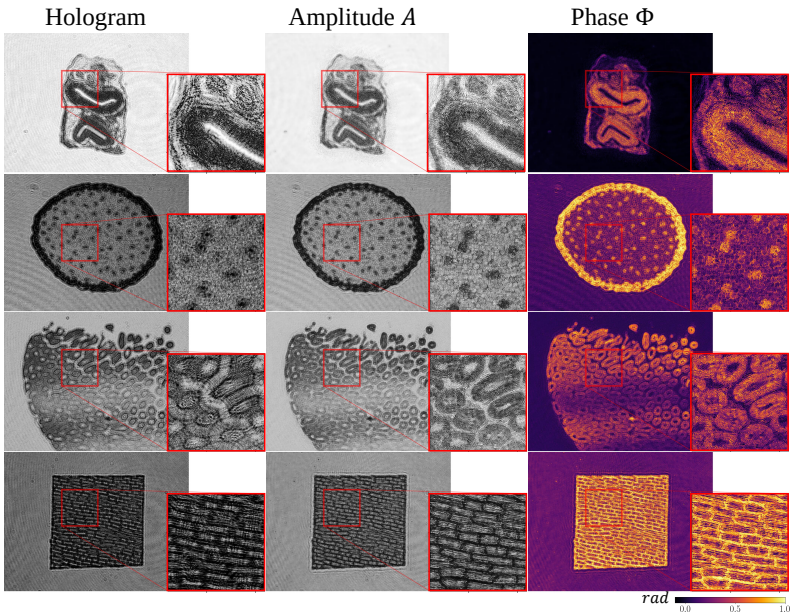


FIGURE 3.21: Reconstructed phase and amplitude distributions from a single input hologram of real samples from [Che+23b] captured with an extended object-detector distance of 5.5 mm.

acceptable results on real data as shown in figure 3.20. Compared to other model-based snapshot solvers shown previously in figure 3.15, it is also possible to see that with larger number of input holograms, results get better and undesirable diffraction artifacts are gradually suppressed.

figure 3.21 shows the reconstructed phase and amplitude images using real samples from [Che+23b] captured at a distance of 5.5 mm from the sensor well beyond the refocusing range used to train HoloADMM (0.5 mm to 1 mm). The proposed model is able to reconstruct both amplitude and high contrast phase images from a single input hologram with a very large refocusing distance. Artifacts in the reconstructed images are mainly due to the lack of input information in this case where $N = 1$.

Furthermore, it is interesting to quantify model robustness to inaccuracies in the refocus distance estimation during inference: although the piezo actuators are very accurate ($\pm$25nm), issues may arise from the estimation algorithm (where gradient sparsity is used as a refocus metric). figure 3.22 shows the model accuracy (i.e., the PSNR of the reconstructed phase) with a fairly high z value deviation starting from $2\delta z = \pm 20\mu$m: the model preserves a good PSNR across all perturbations. Note no large deviation in the distance estimation across different sample types were observed.

3.5 Concluding Remarks

In this second part of the thesis HoloADMM is introduced, an approach that combines interpretability of model-based solvers and the large learning capacity of

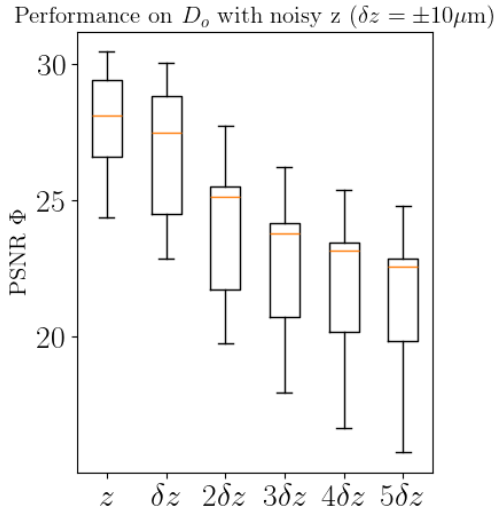


FIGURE 3.22: Reconstruction quality in terms of PSNR with inaccurate estimates of the refocusing distance.

deep neural networks. It is demonstrated that by leveraging large-scale synthetic datasets, high-quality phase imaging capability can be achieved. Additionally, the proposed approach exhibits strong generalization abilities, seamlessly extending to real captured holographic data with notable accuracy. This not only highlights the promise of this methodology but also suggests its potential for practical applications across diverse domains, from biomedical imaging to materials science and beyond. Nonetheless, there are other avenues for potential investigation, such as joint computational refocusing. This is a natural extension of this work and will be further investigated.

Model Limitations HoloADMM relies on an accurate estimation of the refocusing distance, since significant deviations in the measurements can result in sub-optimal reconstructions with real data. Moreover, the complete absence of amplitude information across the entire scene can impact the registration module and subsequently affect the spatial quality of the reconstruction. Lastly, there are instances where the model excessively enhances image contrast in areas where it shouldn't, leading to erroneous phase shift values in those regions. This phenomenon occurs due to the presence of stray light, which provides misleading information for the network that inadvertently amplify its effect in the reconstructed phase image. One could address the above limitations by: 1) Introducing iterative refinement steps that gradually adjust the refocusing distance based on feedback from the reconstruction quality metrics. 2) Use complex edge magnitude to evaluate focus maps. 3) Incorporate a stray light correction module that pre-processes the holographic data to remove or mitigate the effects of stray light.

Chapter 4

Conclusions

4.1 Summary of Findings

Forward steps in the field of computational image sensing have been made by this work, particularly in the areas of spectral imaging and holographic phase imaging. The primary focus was on developing techniques that address the inherent challenges in these advanced imaging methods, particularly the inverse problems that arise during image reconstruction.

The research presented in this thesis dealt with computational reconstruction approaches for spectral imaging and phase imaging where learning-based solution have been proposed to tackle the ill-posed inverse nature of the reconstruction problems in the two cases. HSRN and its subsequent variant HSRN+ have been proposed for joint spectral image recovery and spatial super-resolution from CTIS and MACTIS measurements with extensive studies on synthetic as well as real data captured using three different spectrometer prototypes. For the phase imaging problem HoloADMM, a deep unrolled architecture, has been proposed. It surpasses current state-of-the-art in terms of image quality and its capability to transfer to new unseen domains. Extensive investigations have been carried out to validate the model performance on synthetic as well as real holographic data captured by a custom-made DIHM prototype.

In a nutshell, the key findings of this research are:

- Enhanced Spectral Imaging: The proposed methods allowed to successfully improved the resolution and accuracy of spectral images, allowing for more detailed analysis of material properties.
- Improved Phase Imaging: The advancements in holographic phase imaging have enabled more accurate reconstruction of phase information, which is critical for applications in medical imaging and material science.
- Robust Computational Techniques: The integration of computational algorithms with imaging techniques has proven effective in solving inverse problems, leading to more reliable and accurate imaging results.

4.2 Practical Implications

The techniques developed in this thesis have some practical implications as discussed before. However, further applications can benefit from such approaches. In medical imaging, the enhanced spectral and phase imaging methods can lead to better diagnostic tools, enabling more precise detection and characterization of diseases. In material science, these techniques provide deeper insights into the structural and compositional properties of materials, which can drive innovations in manufacturing and quality control.

Furthermore, the computational methods introduced in this thesis are not limited to the specific applications discussed. They can be adapted and extended to other imaging modalities, broadening their impact across various fields of engineering and science.

4.3 Future Research Directions

While this thesis has addressed several key challenges in imaging science, there remain numerous opportunities for further research. Some potential directions for future work include:

- Domain Adaptation: In order to bridge the ever increasing gap between synthetic and real acquisition settings, domain adaptation techniques are needed in order to enhance model performance on unseen real world data while exploiting the availability of synthetic data in the training phase. Such techniques should be further investigated.
- Extension of Imaging Techniques to New Modalities: Exploring the application of the developed methods to other imaging modalities, such as X-ray or ultrasound imaging.
- Real-Time Imaging and Reconstruction: Developing algorithms that can achieve real-time imaging and reconstruction, which would be highly beneficial for dynamic studies in both medical and industrial applications.
- Physics based Learning Integration: Incorporating physical models to further enhance image reconstruction accuracy and constrain the possible solution space, particularly in solving complex inverse problems.
- Optimization for Hardware Implementation: Adapting the proposed methods for efficient implementation on hardware of mobile platforms, making them more accessible for real-world applications.

4.4 Final Remarks

This thesis has contributed to the advancement of computational imaging by developing innovative techniques for spectral and holographic phase imaging. These contributions not only enhance our understanding and capabilities in these specific areas but also lay the groundwork for future advancements in the broader field of image based sensing. The ongoing evolution of imaging technology promises to continue transforming our ability to observe and understand the world, and this thesis is a small step forward in that journey.

Bibliography

[ABS16] Boaz Arad and Ohad Ben-Shahar. "Sparse Recovery of Hyperspectral Signal from Natural RGB Images". In: *European Conference on Computer Vision*. Springer. 2016, pp. 19–34.

[Ahl+22] Mads J Ahlebæk et al. "The hybrid approach–Convolutional Neural Networks and Expectation Maximization Algorithm–for Tomographic Reconstruction of Hyperspectral Images". In: *arXiv preprint arXiv:2205.15772* (2022).

[Ama+23a] Simon Amann et al. "Design and realization of a miniaturized high resolution computed tomography imaging spectrometer". In: *Journal of the European Optical Society-Rapid Publications* 19.2 (2023), p. 34.

[Ama+23b] Simon Amann et al. "Parallelized computed tomography imaging spectrometer". In: *Digital Optical Technologies 2023*. Vol. 12624. SPIE. 2023, pp. 71–77.

[Bäc+23] Paul Bäcker et al. "Detecting Tar Contaminated Samples in Roadrubble using Hyperspectral Imaging and Texture Analysis". In: *OCM 2023-Optical Characterization of Materials: Conference Proceedings*. KIT Scientific Publishing. 2023, p. 11.

[Bay76] Bryce E Bayer. "Color imaging array". In: *United States Patent 3,971,065* (1976).

[BDF07] José M Bioucas-Dias and Mário AT Figueiredo. "A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration". In: *IEEE Transactions on Image processing* 16.12 (2007), pp. 2992–3004.

[Bil01] Griff Bilbro. *Technology options for multi-spectral infrared cameras*. Tech. rep. NORTH CAROLINA STATE UNIV AT RALEIGH DEPT OF ELECTRICAL and COMPUTER ENGINEERING, 2001.

[BLH23] Jürgen Beyerer, Thomas Längle, and Michael Heizmann, eds. *OCM 2023 - Optical Characterization of Materials : Conference Proceedings*. 6th International Conference on Optical Characterization of Materials. OCM 2023 (Karlsruhe, Deutschland, Mar. 22–23, 2023). 2023. 178 pp. ISBN: 978-3-7315-1274-5. DOI: 10.5445/KSP/1000155014.

[Boy+11] Stephen Boyd et al. "Distributed optimization and statistical learning via the alternating direction method of multipliers". In: *Foundations and Trends® in Machine learning* 3.1 (2011), pp. 1–122.

[Bru+06] Nicola Brusco et al. "A system for 3D modeling frescoed historical buildings with multispectral texture information". In: *Machine Vision and Applications* 17.6 (2006), pp. 373–393.

[BS42] CR Burch and JPP Stock. "Phase-contrast microscopy". In: *Journal of Scientific Instruments* 19.5 (1942), p. 71.

[BSA10] Johannes Brauers, Claude Seiler, and Til Aach. "Direct PSF estimation using a random noise target". In: *Digital Photography VI*. Vol. 7537. SPIE. 2010, pp. 96–105.

[BV92] Theodor V Bulygin and Gennady N Vishnyakov. "Spectrotomography: a new method of obtaining spectrograms of two-dimensional objects". In: *Analytical Methods for Optical Tomography*. Vol. 1843. SPIE. 1992, pp. 315–322.

[Che+22] Hanlong Chen et al. "Fourier Imager Network (FIN): A deep neural network for hologram reconstruction with superior external generalization". In: *Light: Science & Applications* 11.1 (2022), p. 254.

[Che+23a] Hanlong Chen et al. "eFIN: Enhanced Fourier Imager Network for generalizable autofocusing and pixel super-resolution in holographic imaging". In: *IEEE Journal of Selected Topics in Quantum Electronics* 29.4: Biophotonics (2023), pp. 1–10.

[Che+23b] Xiwen Chen et al. "DH-GAN: a physics-driven untrained generative adversarial network for holographic imaging". In: *Optics Express* 31.6 (2023), pp. 10114–10135.

[Cin+17] Patricia Cintora et al. "Cell density modulates intracellular mass transport in neural networks". In: *Cytometry Part A* 91.5 (2017), pp. 503–509.

[Cou60] Georges Courtes. "Méthodes d'observation et étude de l'hydrogène interstellaire en émission". In: *Annales d'Astrophysique*. Vol. 23. 1960, p. 115.

[CWE16] Stanley H Chan, Xiran Wang, and Omar A Elgendy. "Plug-and-play ADMM for image restoration: Fixed-point convergence and applications". In: *IEEE Transactions on Computational Imaging* 3.1 (2016), pp. 84–98.

[CWH22] Ni Chen, Congli Wang, and Wolfgang Heidrich. "Differentiable holography". In: (2022). DOI: 10.1002/lpor.202200828.

[Dab+07] Kostadin Dabov et al. "Image denoising by sparse 3-D transform-domain collaborative filtering". In: *IEEE Transactions on Image Processing* 16.8 (2007), pp. 2080–2095.

[Dab+08] Kostadin Dabov et al. "Image restoration by sparse 3D transform-domain collaborative filtering". In: *Image Processing: Algorithms and Systems VI*. Vol. 6812. SPIE. 2008, pp. 62–73.

[DDDM04] Ingrid Daubechies, Michel Defrise, and Christine De Mol. "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint". In: *Communications on Pure and Applied Mathematics* 57.11 (2004), pp. 1413–1457.

[Den+09a] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[Den+09b] Loïc Denis et al. "Inline hologram reconstruction with sparsity constraints". In: *Optics letters* 34.22 (2009), pp. 3475–3477.

[Dia+20] Foivos I Diakogiannis et al. "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 162 (2020), pp. 94–114.

[Dja+23] Ervienatasia Djaw et al. "Method Development for Spatially Resolved Detection of Adulterated Minced Meat". In: *OCM 2023-Optical Characterization of Materials: Conference Proceedings*. KIT Scientific Publishing. 2023, p. 65.

[Dou+20] Clément Douarre et al. "On the value of CTIS imagery for neural-network-based classification: a simulation perspective". In: *Applied optics* 59.28 (2020), pp. 8697–8710.

[Dou+21] Clément Douarre et al. "CTIS-Net: a neural network architecture for compressed learning based on computed tomography imaging spectrometers". In: *IEEE Transactions on Computational Imaging* 7 (2021), pp. 572–583.

[DRS20] Jiangxin Dong, Stefan Roth, and Bernt Schiele. "Deep wiener deconvolution: Wiener meets deep learning for image deblurring". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1048–1059.

[Dud+22] Akshay Dudhane et al. "Burst image restoration and enhancement". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5759–5768.

[EA+22] Wassim A El Ahmar et al. "Multiple Object Detection and Tracking in the Thermal Spectrum". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 277–285.

[Écl31] Commission Internationale de l Éclairage. *Commission internationale de l eclairage proceedings*. 1931.

[EPF14] David Eigen, Christian Puhrsch, and Rob Fergus. "Depth map prediction from a single image using a multi-scale deep network". In: *Advances in neural information processing systems* 27 (2014).

[Fie78] James R Fienup. "Reconstruction of an object from the modulus of its Fourier transform". In: *Optics letters* 3.1 (1978), pp. 27–29.

[Foi+08] Alessandro Foi et al. "Practical Poissonian-Gaussian noise modeling and fitting for single-image raw-data". In: *IEEE transactions on image processing* 17.10 (2008), pp. 1737–1754.

[Gab49] Dennis Gabor. "Microscopy by reconstructed wave-fronts". In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 197.1051 (1949), pp. 454–487.

[GBI09] Daniel Glasner, Shai Bagon, and Michal Irani. "Super-resolution from a single image". In: *2009 IEEE 12th international conference on computer vision*. IEEE. 2009, pp. 349–356.

[Geh+07] Michael E Gehm et al. "Single-shot compressive spectral imaging with a dual-disperser architecture". In: *Optics express* 15.21 (2007), pp. 14013–14027.

[GM76] Daniel Gabay and Bertrand Mercier. "A dual algorithm for the solution of nonlinear variational problems via finite element approximation". In: *Computers & mathematics with applications* 2.1 (1976), pp. 17–40.

[God+16] T. M. Godden et al. "Phase calibration target for quantitative phase imaging with ptychography". In: *Opt. Express* 24.7 (2016), pp. 7679–7692.

[Goo05] Joseph W Goodman. *Introduction to Fourier optics*. Roberts and Company Publishers, 2005.

[GS94] RW Gerchberg and WO Saxton. "A practical algorithm for the determination of phase from image and diffraction plane pictures". In: *SPIE milestone series MS* 93 (1994), pp. 306–306.

[GSTF08] Manuel Guizar-Sicairos, Samuel T Thurman, and James R Fienup. "Efficient subpixel image registration algorithms". In: *Optics letters* 33.2 (2008), pp. 156–158.

[Gub16] Nitzan Guberman. "On complex valued convolutional neural networks". In: *arXiv preprint arXiv:1602.09046* (2016).

[GY95] Donald Geman and Chengda Yang. "Nonlinear image recovery with half-quadratic regularization". In: *IEEE transactions on Image Processing* 4.7 (1995), pp. 932–946.

[HB81] H Harrison and Swindell W Barrett. *Radiological imaging: the theory of image, formation, detection, and processing*. 1981.

[HDS07] Nathan Hagen, Eustace L Dereniak, and David T Sass. "Fourier methods of improving reconstruction speed for CTIS imaging spectrometers". In: *Imaging Spectrometry XII*. Vol. 6661. SPIE. 2007, pp. 15–25.

[HK13] Nathan A Hagen and Michael W Kudenov. "Review of snapshot spectral imaging technologies". In: *Optical Engineering* 52.9 (2013), p. 090901.

[HKW12] Ralf Habel, Michael Kudenov, and Michael Wimmer. "Practical spectral photography". In: *Computer graphics forum*. Vol. 31. 2pt2. Wiley Online Library. 2012, pp. 449–458.

[HSA15] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. "Single image super-resolution from transformed self-exemplars". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 5197–5206.

[HSB02] Jon Yngve Hardeberg, Francis JM Schmitt, and Hans Brettel. "Multispectral color image capture using a liquid crystal tunable filter". In: *Optical engineering* 41.10 (2002), pp. 2532–2548.

[Hua+22] Wei-Chih Huang et al. "The application of convolutional neural networks for tomographic reconstruction of hyperspectral images". In: *Displays* 74 (2022), p. 102218.

[Hua+23] Luzhe Huang et al. "Self-supervised learning of hologram reconstruction using physics consistency". In: *Nature Machine Intelligence* 5.8 (2023), pp. 895–907.

[HWC21] Weizhe Han, Qianlong Wang, and Weiwei Cai. "Computed tomography imaging spectrometry based on superiorization and guided image filtering". In: *Optics Letters* 46.9 (2021), pp. 2208–2211.

[Jha+19] Debesh Jha et al. "Resunet++: An advanced architecture for medical image segmentation". In: *2019 IEEE international symposium on multimedia (ISM)*. IEEE. 2019, pp. 225–2255.

[Jia+17] Yan Jia et al. "From RGB to spectrum for natural scenes via manifold-based mapping". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 4705–4713.

[JSK08] Neel Joshi, Richard Szeliski, and David J Kriegman. "PSF estimation using sharp edge prediction". In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2008, pp. 1–8.

[KB14] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[Kha+18] Muhammad Jaleed Khan et al. "Modern trends in hyperspectral image analysis: A review". In: *Ieee Access* 6 (2018), pp. 14118–14129.

[Kit+10] David Kittle et al. “Multiframe image estimation for coded aperture snapshot spectral imagers”. In: *Applied optics* 49.36 (2010), pp. 6824–6833.

[KVB08] Björn Kemper and Gert Von Bally. “Digital holographic microscopy for live cell applications and technical inspection”. In: *Applied optics* 47.4 (2008), A52–A61.

[Led+17] Christian Ledig et al. “Photo-realistic single image super-resolution using a generative adversarial network”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 4681–4690.

[LF07] Tatiana Latychevskaia and Hans-Werner Fink. “Solution to the twin image problem in holography”. In: *Physical review letters* 98.23 (2007), p. 233901.

[Li+10] Xuelong Li et al. “A multi-frame image super-resolution method”. In: *Signal Processing* 90.2 (2010), pp. 405–414.

[Li+18] Qifeng Li et al. “A low-rank estimation method for CTIS image reconstruction”. In: *Measurement Science and Technology* 29.9 (2018), p. 095401.

[Li+19] Sheng Li et al. “Fast spatio-temporal residual network for video super-resolution”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 10522–10531.

[Li+20] Zongyi Li et al. “Fourier neural operator for parametric partial differential equations”. In: *arXiv preprint arXiv:2010.08895* (2020).

[Lim+17] Bee Lim et al. “Enhanced deep residual networks for single image super-resolution”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*. 2017, pp. 136–144.

[Liu+18] Yang Liu et al. “Rank minimization for snapshot compressive imaging”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.12 (2018), pp. 2990–3006.

[LTO12] Xinhao Liu, Masayuki Tanaka, and Masatoshi Okutomi. “Noise level estimation using weak textured patches of a single noisy image”. In: *2012 19th IEEE International Conference on Image Processing*. IEEE. 2012, pp. 665–668.

[Mak+22] Maksim Makarenko et al. “Real-time hyperspectral imaging in hardware via trained metasurface encoders”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 12692–12702.

[Mel+24] Mazen Mel et al. “HoloADMM: High-Quality Holographic Complex Field Recovery”. In: *European Conference on Computer Vision (ECCV)*. 2024.

[Men+21] Ziyi Meng et al. "Self-supervised neural networks for spectral snapshot compressive imaging". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 2622–2631.

[MGZ22] Mazen Mel, Alexander Gatto, and Pietro Zanuttigh. "Joint Reconstruction and Super Resolution of Hyper-Spectral CTIS Images". In: *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. URL: https://bmvc2022.mpi-inf.mpg.de/1063.pdf.

[MGZ24] Mazen Mel, Alexander Gatto, and Pietro Zanuttigh. "Joint Reconstruction and Spatial Super-resolution of Hyper-Spectral CTIS Images via Multi-Scale Refinement". In: *IEEE transactions on Computational Imaging* (2024).

[MH12] Ajmal Mian and Richard Hartley. "Hyperspectral video restoration using optical flow and sparse coding". In: *Optics express* 20.10 (2012), pp. 10658–10673.

[Mir+12] Mustafa Mir et al. "Quantitative phase imaging". In: *Progress in optics* 57.133-37 (2012), p. 217.

[Mom+19] Fabien Momey et al. "From Fienup's phase retrieval techniques to regularized inversion for in-line holography: tutorial". In: *JOSA A* 36.12 (2019), pp. D62–D80.

[Mon+15] Yusukex Monno et al. "A practical one-shot multispectral imaging system using a single image sensor". In: *IEEE Transactions on Image Processing* 24.10 (2015), pp. 3048–3059.

[Mos+15] Ali Mosleh et al. "Camera intrinsic blur kernel estimation: A reliable framework". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 4961–4968.

[MT95] Ali Seif A Mshimba and Wolfgang Tutschke. *Functional analytic methods in complex analysis and applications to partial differential equations*. World Scientific, 1995.

[MTZM18] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. "The contextual loss for image transformation with non-aligned data". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 768–783.

[NQL21] Farhad Niknam, Hamed Qazvini, and Hamid Latifi. "Holographic optical field recovery using a regularized untrained deep decoder network". In: *Scientific reports* 11.1 (2021), p. 10903.

[OY91] Takayuki Okamoto and Ichirou Yamaguchi. "Simultaneous acquisition of spectral image information". In: *Optics letters* 16.16 (1991), pp. 1277–1279.

[Par+08] YongKeun Park et al. "Refractive index maps and membrane dynamics of human red blood cells parasitized by Plasmodium falciparum". In: *Proceedings of the National Academy of Sciences* 105.37 (2008), pp. 13730–13735.

[PDP18] YongKeun Park, Christian Depeursinge, and Gabriel Popescu. "Quantitative phase imaging in biomedicine". In: *Nature photonics* 12.10 (2018), pp. 578–589.

[PE87] Wallace M Porter and Harry T Enmark. "A system overview of the airborne visible/infrared imaging spectrometer (AVIRIS)". In: *Imaging Spectroscopy II*. Vol. 834. SPIE. 1987, pp. 22–31.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer. 2015, pp. 234–241.

[Riv+18] Yair Rivenson et al. "Phase recovery and holographic image reconstruction using deep learning in neural networks". In: *Light: Science & Applications* 7.2 (2018), pp. 17141–17141.

[RL71] GN Ramachandran and AV Lakshminarayanan. "Three-dimensional reconstruction from radiographs and electron micrographs: application of convolutions instead of Fourier transforms". In: *Proceedings of the National Academy of Sciences* 68.9 (1971), pp. 2236–2240.

[ROF92] Leonid I Rudin, Stanley Osher, and Emad Fatemi. "Nonlinear total variation based noise removal algorithms". In: *Physica D: nonlinear phenomena* 60.1-4 (1992), pp. 259–268.

[Rom+22] Robin Rombach et al. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.

[RXL19] Zhenbo Ren, Zhimin Xu, and Edmund Y Lam. "End-to-end deep learning framework for digital holographic reconstruction". In: *Advanced Photonics* 1.1 (2019), pp. 016004–016004.

[Shi+16] Wenzhe Shi et al. "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 1874–1883.

[Sim+21] Adriano Simonetto et al. "Semi-supervised Deep Learning Techniques for Spectrum Reconstruction". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 7767–7774.

[SK24] H. Paul Urbach Sander Konijnenberg Aurèle J.L. Adam. *BSc Optics: 2nd edition*. TUDelft, 2024. DOI: https://doi.org/10.59490/tb.91.

[SZ14] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[Tam+17] Miu Tamamitsu et al. "Comparison of Gini index and Tamura coefficient for holographic autofocusing based on the edge sparsity of the complex optical wavefront". In: *arXiv preprint arXiv:1708.08055* (2017).

[Tek+13] Mustafa Teke et al. "A short survey of hyperspectral remote sensing applications in agriculture". In: *2013 6th international conference on recent advances in space technologies (RAST)*. IEEE. 2013, pp. 171–176.

[Vol00] Curtis Earl Volin. "Portable snapshot infrared imaging spectrometer". PhD thesis. The University of Arizona, 2000.

[VS70] CM Vest and DW Sweeney. "Holographic interferometry of transparent objects with illumination derived from phase gratings". In: *Applied Optics* 9.10 (1970), pp. 2321–2325.

[Wag+08] Ashwin Wagadarikar et al. "Single disperser design for coded aperture snapshot spectral imaging". In: *Applied optics* 47.10 (2008), B44–B51.

[Wan+18] Xintao Wang et al. "Esrgan: Enhanced super-resolution generative adversarial networks". In: *Proceedings of the European conference on computer vision (ECCV) workshops*. 2018, pp. 0–0.

[WBH20] Larz White, W Bryan Bell, and Ryan Haygood. "Accelerating computed tomographic imaging spectrometer reconstruction using a parallel algorithm exploiting spatial shift-invariance". In: *Optical Engineering* 59.5 (2020), p. 055110.

[WC23] Luoxiang Wu and Weiwei Cai. "CTIS-GAN: computed tomography imaging spectrometry based on a generative adversarial network". In: *Applied Optics* 62.10 (2023), pp. 2422–2433.

[Wro+19] Bartlomiej Wronski et al. "Handheld multi-frame super-resolution". In: *ACM Transactions on Graphics (TOG)* 38.4 (2019), pp. 1–18.

[Wu+21] Yufeng Wu et al. "Dense-U-net: dense encoder–decoder network for holographic imaging of 3D particle fields". In: *Optics Communications* 493 (2021), p. 126970.

[Xiw+24] Chen Xiwen et al. "Enhancing digital hologram reconstruction using reverse-attention loss for untrained physics-driven deep learning models with uncertain distance". In: *SPIE Photonics West 2024*. 2024.

[Yan+10] Jianchao Yang et al. "Image super-resolution via sparse representation". In: *IEEE transactions on image processing* 19.11 (2010), pp. 2861–2873.

[Yas+10] Fumihito Yasuma et al. “Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum”. In: *IEEE Transactions on Image Processing* 19.9 (2010), pp. 2241–2253.

[Yua16] Xin Yuan. “Generalized alternating projection based total variation minimization for compressive sensing”. In: *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2016, pp. 2539–2543.

[Yua+23] Lei Yuan et al. “Super-resolution computed tomography imaging spectrometry”. In: *Photonics research* 11.2 (2023), pp. 212–224.

[YWL13] Jiangye Yuan, DeLiang Wang, and Rongxing Li. “Remote sensing image segmentation by combining spectral and texture features”. In: *IEEE Transactions on geoscience and remote sensing* 52.1 (2013), pp. 16–24.

[Zei+10] Matthew D Zeiler et al. “Deconvolutional networks”. In: *2010 IEEE Computer Society Conference on computer vision and pattern recognition*. IEEE. 2010, pp. 2528–2535.

[Zha+17] Kai Zhang et al. “Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising”. In: *IEEE Transactions on Image Processing* 26.7 (2017), pp. 3142–3155.

[Zha+18] Wenhui Zhang et al. “Twin-image-free holography: a compressive sensing approach”. In: *Physical review letters* 121.9 (2018), p. 093902.

[Zha+21] Shipeng Zhang et al. “Learning tensor low-rank prior for hyperspectral image reconstruction”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 12006–12015.

[Zhe+21] Siming Zheng et al. “Deep plug-and-play priors for spectral snapshot compressive imaging”. In: *Photonics Research* 9.2 (2021), B18–B29.

[Zie+19] Michał Ziemczonok et al. “3D-printed biological cell phantom for testing 3D quantitative phase imaging systems”. In: *Scientific reports* 9.1 (2019), p. 18872.

[Zim+22] Markus Zimmermann et al. “Deep learning-based hyperspectral image reconstruction from emulated and real computed tomography imaging spectrometer data”. In: *Optical Engineering* 61.5 (2022), p. 053103.

[ZLW18] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. “Road extraction by deep residual u-net”. In: *IEEE Geoscience and Remote Sensing Letters* 15.5 (2018), pp. 749–753.