Deep Learning based Distillation Algorithm for Designing a Highly Physically Constrained Computational Imaging System

León Santiago Suárez Rodríguez

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Systems Engineering and Informatics

Advisor

Henry Arguello Fuentes

PhD. in Electrical and Computer Engineering

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Ingeniería de Sistemas e Informática

Maestría en Ingeniería de Sistemas e Informática

2025

Dedicatory

To my family

Acknowledgments

I would like to express my sincere gratitude to the following individuals and groups for their invaluable support throughout this journey:

- **Professor Henry Arguello** – For his exceptional guidance, mentorship, and support as my advisor.
- **Roman Jacome and Paul Goyes** – For their insightful mentorship and encouragement.
- **Luis and Romario ("The Three Musketeers")** – For their camaraderie and unwavering support.
- **Romario, Roman, Pablo and Emmanuel ("The Kamikaze Team")** – For their dedication and shared efforts in our collaborations.
- **Paul, Ana, Luis, and Javier ("The Geophysics Team")** – For their friendship, teamwork, and motivation throughout our projects.
- **My family** – For their unconditional support, patience, and encouragement throughout this journey.
- **David, Brayan, Roman, Romario, and Emmanuel ("The Colibrí Team")** – For enriching my knowledge and fostering a collaborative environment.

- **The HDSP group** – For their companionship, support, and collaboration throughout this process.

Table of Contents

List of Figures

List of Tables

Abstract

Title: Deep Learning based Distillation Algorithm for Designing a Highly Physically Constrained Computational Imaging System [*]

Autor: León Santiago Suárez Rodríguez [**]

Keywords: Knowledge distillation, computational imaging systems, end-to-end optimization, magnetic resonance imaging, coded aperture systems.

Description: Computational imaging (CI) systems extend the capabilities of traditional imaging systems by encoding high-dimensional signal information into low-dimensional coded projections, which are subsequently decoded using computational algorithms. The design of the physical encoder is crucial for accurate image reconstruction, as it determines how the scene is sampled and encoded, directly affecting the quality and quantity of the encoded information and its subsequent reconstruction. Currently, CI systems are designed using an end-to-end (E2E) optimization approach, where the encoder is represented as a neural network layer and is jointly optimized with the computational decoder. However, the performance of E2E optimization is significantly reduced by the physical constraints imposed on the encoder, such as binarization, light throughput, and the compression ratio. Moreover, since E2E learns the parameters of the encoder by backpropagating the reconstruction error, it does not promote optimal intermediate outputs and suffers from gradient vanishing.

To address these limitations, this research reinterprets the concept of knowledge distillation—traditionally used to train smaller neural networks by transferring knowledge from a larger pretrained model—to design

---

[*] Master Thesis

[**] Faculty of Physicomechanical. Department of Systems Engineering. Advisor: Henry Arguello Fuentes, Ph.D. in Electrical and Computer Engineering.

a physically constrained CI system by transferring knowledge from a pretrained, less-constrained CI system. The proposed approach involves three steps: First, given the original CI system (student), a teacher system is created by relaxing the constraints on the student's encoder. Second, the teacher is optimized to solve a less-constrained version of the student's problem. Third, the teacher guides the training of the highly constrained student through two proposed knowledge transfer functions, targeting both the encoder's structure and the decoder feature space.

This approach was validated on three representative CI modalities: magnetic resonance, single-pixel, and compressive spectral imaging. Simulations demonstrate that a teacher system with an encoder structure similar to the student's—having a comparable number of measurements or similar nature of the codification basis—provides effective guidance. This leads to significantly improved reconstruction performance and encoder design for the student, outperforming both E2E optimization and traditional non-data-driven encoder designs.

Resumen

Título: Algoritmo de Destilación basado en Aprendizaje Profundo para el Diseño de un Sistema de Imagen Computacional Altamente Restringido Físicamente [*]

Autor: León Santiago Suárez Rodríguez [**]

Palabras Clave: Destilación de conocimiento, sistemas de imagen computacional, optimización de extremo a extremo, imagen por resonancia magnética, sistemas de apertura codificada.

Descripción: Los sistemas de adquisición de imágenes computacionales (CI, por sus siglas en inglés) extienden las capacidades de los sistemas de imágenes tradicionales al codificar información de señales de alta dimensión en proyecciones codificadas de baja dimensión, que posteriormente se decodifican mediante algoritmos computacionales. El diseño del codificador físico es crucial para una reconstrucción precisa de la imagen, ya que determina cómo se muestrea y codifica la escena, afectando directamente la calidad y cantidad de la información codificada y su posterior reconstrucción. Actualmente, los sistemas CI se diseñan mediante un enfoque de optimización de extremo a extremo (E2E), donde el codificador se representa como una capa de red neuronal y se optimiza conjuntamente con el decodificador computacional. Sin embargo, el rendimiento de la optimización E2E se ve significativamente reducido por las restricciones físicas impuestas al codificador, como la binarización, la transmisión de luz y la relación de compresión. Además, dado que E2E aprende los parámetros del codificador retropropagando el error de reconstrucción, no promueve salidas intermedias óptimas y sufre del problema de desaparición del gradiente.

Para abordar estas limitaciones, esta investigación reinterpreta el concepto de destilación de conocimien-

---

[*] Trabajo de grado

[**] Facultad de Ingenierias Fisicomecánicas. Escuela de Ingenieria de Sistemas e Informatica. Director: Henry Arguello Fuentes, Doctorado en Ingeniería Eléctrica y Computación.

to—tradicionalmente utilizado para entrenar redes neuronales más pequeñas transfiriendo conocimiento desde un modelo preentrenado más grande—para diseñar un sistema CI físicamente restringido transfiriendo conocimiento desde un sistema CI preentrenado con menos restricciones. El enfoque propuesto implica tres pasos: Primero, dado el sistema CI original (estudiante), se crea un sistema maestro relajando las restricciones en el codificador del estudiante. Segundo, el maestro se optimiza para resolver una versión menos restringida del problema del estudiante. Tercero, el maestro guía el entrenamiento del estudiante altamente restringido mediante dos funciones de transferencia de conocimiento propuestas, dirigidas tanto a la estructura del codificador como al espacio de características del decodificador.

Este enfoque fue validado en tres modalidades representativas de CI: resonancia magnética, imagen de un solo píxel e imagen espectral compresiva. Las simulaciones demuestran que un sistema maestro con una estructura de codificador similar a la del estudiante—con un número comparable de mediciones o una naturaleza similar en la base de codificación—proporciona una guía efectiva. Esto conduce a una mejora significativa en el rendimiento de reconstrucción y en el diseño del codificador del estudiante, superando tanto la optimización E2E como los diseños de codificadores tradicionales no basados en datos.

Research Products

Contributions of the thesis

- This research introduces a novel knowledge distillation framework for computational imaging system design, addressing the performance limitations imposed by physical constraints on the encoder in traditional end-to-end optimization. In this approach, the constraints on the encoder in the highly constrained student are relaxed to create the teacher, which is then trained and used to guide the student's learning.
- Two knowledge transfer loss functions are proposed to facilitate knowledge transfer from the teacher to the student: (1) encoder loss, which aligns the student's encoder structure with that of the teacher, and (2) decoder loss, which aligns the feature spaces of their respective computational decoders.
- This research demonstrates that teachers with a similar number of measurements and an encoder with a similar codification basis to that of the student provide effective guidance, leading to significant improvements in both the student's encoder design and its reconstruction performance.
- The proposed approach was validated across three widely used computational imaging systems—magnetic resonance imaging, the single-pixel camera, and the single-disperser coded aperture snapshot spectral imager—showing superior performance compared to traditional E2E optimization and non-data-driven encoder designs.

Publications

The developments of this thesis have been disseminated in several international journals and conferences.

Journal papers:

1. P. Goyes-Peñafiel, **L. Suárez-Rodríguez**, C. V. Correa and H. Arguello, "GAN Supervised Seismic Data Reconstruction: An Enhanced Learning for Improved Generalization," in IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1-10, 2024, Art no. 5921910, doi: 10.1109/TGRS.2024.3434474.

2. **L. Suárez-Rodríguez**, R. Jacome and H. Arguello, "Distilling Knowledge for Designing Computational Imaging Systems," in IEEE Transactions on Computational Imaging (Under review).

Conference papers:

1. **L. Suarez-Rodriguez**, R. Jacome and H. Arguello, "Highly Constrained Coded Aperture Imaging Systems Design Via a Knowledge Distillation Approach," 2024 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 2024, pp. 3993-3999, doi: 10.1109/ICIP51287.2024.10648100.

2. R. Jacome, **L. Suarez**, R. Gualdrón-Hurtado, L. Gonzalez and H. Arguello, "Learning to Reconstruct Signals With Inexact Sensing Operator via Knowledge Distillation," ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal

Processing (ICASSP), Hyderabad, India, 2025, pp. 1-5,
doi: 10.1109/ICASSP49660.2025.10887652.

3. E. Martinez, **L. Suarez**, R. Gualdrón-Hurtado, R. Jacome and H. Arguello, "Compressive Imaging Reconstruction via Conditional Diffusion Model With Augmented Measurements" ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1-5,
doi: 10.1109/ICASSP49660.2025.10889114.

4. R. Gualdrón-Hurtado, R. Jacome, **L. Suarez**, E. Martinez and H. Arguello, "Improving Compressive Imaging Recovery via Measurement Augmentation," ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1-5, doi: 10.1109/ICASSP49660.2025.10888734.

1. Introduction

Computational imaging (CI) systems have extended the capabilities of traditional imaging systems by integrating computational algorithms with physical image acquisition, surpassing conventional imaging limitations such as blurring, dynamic range, spatial resolution and depth of field Bhandari et al. (2022); Liang (2024). CI systems employ a physical encoder to encode the desired high-dimensional signal information into low-dimensional coded projections, which are then reconstructed by a computational decoder Arguello et al. (2023). CI systems have been applied in several imaging fields such as optical imaging, seismic imaging, computational microscopy, and medical imaging Bhandari et al. (2022); Karl et al. (2023); Wetzstein et al. (2020). The performance of these systems relies on the effective design of the encoder, as it determines how the scene is sampled and encoded, affecting the quality and amount of information that can be encoded and subsequently reconstructed by the computational decoder.

Traditionally, CI system design has relied on predefined random or structured patterns, which have proven effective for different imaging modalities. Examples include Hadamard and Fourier coded apertures (CAs) Zhang et al. (2017); Duarte et al. (2008), radial and spiral undersampling patterns for magnetic resonance imaging (MRI) Geethanath et al. (2013), and regular or jittered sampling strategies in seismic survey acquisition Hernandez-Rojas and Arguello (2024). Similarly, diffractive optical elements (DOEs) have been designed using Fresnel Miyamoto (1961), saddle Bernet (2018), and spiral Oemrawsingh et al. (2004)

lenses, while 3D ultrasound imaging has leveraged 2D sparse arrays based on deterministic or stochastic spiral distributions Ramalli et al. (2022). Furthermore, analytical designs leveraging compressive sensing principles have been widely explored across various applications Correa et al. (2016); Henry and Gonzalo (2011); Lustig et al. (2007); Vasanawala et al. (2011); Mosher et al. (2014, 2017). While these traditional designs have been effective in their respective contexts, they are typically limited by hand-crafted signal properties, limiting their performance.

In contrast, current approaches optimize CI systems in a data-driven manner using an end-to-end (E2E) framework, enabling high-performance task-specific CI systems. This method jointly optimizes the encoder and computational decoder for a specific imaging task, incorporating deep learning by representing the image formation model as a neural network layer, with subsequent layers acting as the computational decoder Arguello et al. (2023). Examples include the design of a DOE for high dynamic range imaging Metzler et al. (2020), the design of a colored CA for depth estimation Lopez et al. (2024), undersampling mask optimization for MRI reconstruction Bahadir et al. (2020), seismic survey design for seismic imaging reconstruction Hernandez-Rojas and Arguello (2024), and microscopic illumination pattern design for image reconstruction Kellman et al. (2019).

The design of the encoder must account for physical constraints. In MRI, for instance, long scans can lead to patient discomfort, motion artifacts, high costs, and long waiting times that may extend for months Bahadir et al. (2020); GharehMohammadi and Sebro (2024); Zaitsev et al. (2015). To address these issues, $k-$space measurements are undersampled to

reduce acquisition time, resulting in a trade-off between obtaining high-quality images and minimizing scan times.

Similarly, optical imaging systems employ CAs to reduce the amount of information required during acquisition while maintaining high-quality image reconstruction. Binary-valued CAs are commonly used because they are easier to fabricate and calibrate than real-valued or quantized CAs, use less integration time, and have lower storage requirements Bacca et al. (2021). However, each element of the binary-valued CA is limited to acquiring only two states: blocking or passing light. Compared to other CA types, this limitation reduces the amount of information that can be encoded and subsequently reconstructed, resulting in less detailed or lower-quality images Arguello et al. (2023); Bacca et al. (2021). Furthermore, CA imaging systems are constrained by the need to acquire a limited number of snapshots to reduce acquisition and processing times, resulting in a trade-off between image quality and the time needed for acquisition and processing. Therefore, these systems aim to minimize the number of acquisitions and maximize the performance on the imaging task Arguello and Arce (2014); Bacca et al. (2021).

When multiple or severe constraints are introduced, CI systems become highly constrained, making the trade-off between practical implementation and reconstruction quality even more critical. For instance, in MRI, higher acceleration factors (AF)(e.g., 8× or 12×) make image recovery increasingly challenging Munoz et al. (2022); Chu et al. (2025), yet they enable faster scans, making the approach practical for real-world implementation. Similarly, in an CA systems, a high compression ratio combined with binary-valued CAs reduces

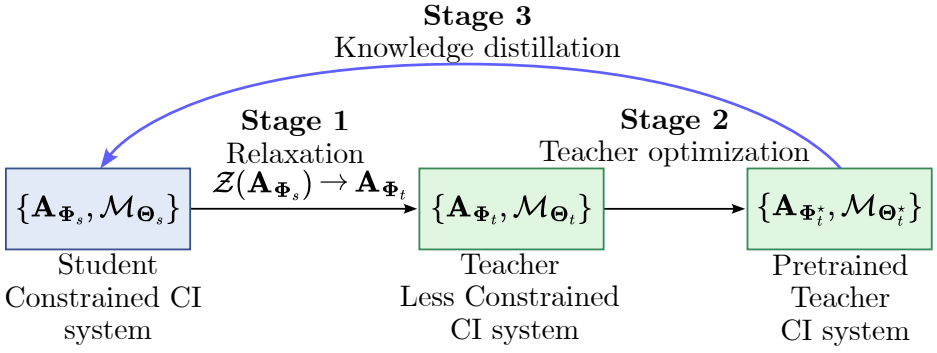
acquisition and processing times and simplifies implementation but increases the difficulty of reconstructing accurate images Arguello et al. (2023).

These constraints of CI systems are incorporated into the E2E optimization problem through regularization functions that constrain the set of optimal values of the imaging formation model parameters Arguello et al. (2023). Consequently, these regularization functions reduce the degrees of freedom of the CI system during training, ultimately degrading its performance on the imaging task Jacome et al. (2023). Furthermore, since the E2E optimization framework learns the encoder parameters by backpropagating the reconstruction error, it does not promote optimal intermediate outputs, as these outputs are unknown. Moreover, the encoder of the CI system may not be optimal as it suffers from the vanishing gradient problem due to being the first layer of the E2E framework Jacome et al. (2023). Thus, it is required to develop new learning techniques to address the implementation constraints of the encoder and find optimality criteria for the design of the encoder.

To address the performance limitations of E2E optimization for CI system design, this work proposes a novel approach inspired by knowledge distillation (KD) Hinton et al. (2015) theory. In KD, a larger and complex neural network, known as the **teacher**, guides the training of a smaller and simpler network, called the **student**, using a knowledge transfer function to effectively convey learned representations and improve performance. The student is specifically designed for resource-constrained environments, and with the teacher's guidance, it can achieve higher performance than it would through standalone training. This research reinterprets the concept of KD for CI system design, where a high-performance,

less-constrained CI system serves as the teacher model, while a more physically constrained CI system acts as the student model. Both CI systems share computational decoders with the same neural network architecture, with the focus on addressing the constraints in the student's encoder. For the simulations, the well-known U-Net architecture was used Ronneberger et al. (2015); however, the proposed methodology can be extended to any computational decoder neural network architecture. The proposed approach follows a three-stage methodology, illustrated in Figure 1: (1) the teacher encoder is created by relaxing the constraints on the student encoder, (2) the teacher encoder is optimized along with its reconstruction decoder to solve a less-constrained version of the student's problem, and (3) the teacher provides guidance to the student by transferring knowledge through various proposed KD loss functions.

Figure 1. The student system is a constrained CI system with encoder $\mathbf{A}_{\mathbf{\Phi}_s}$ and decoder $\mathcal{M}_{\mathbf{\Theta}_s}$. In the first stage, by relaxing the student encoder constraints, a teacher encoder $\mathbf{A}_{\mathbf{\Phi}_t}$ is derived. In the second stage, the teacher encoder and its reconstruction network $\mathcal{M}_{\mathbf{\Theta}_t}$ are optimized to solve a less-constrained version of the student's problem, resulting in $\mathbf{A}_{\mathbf{\Phi}_t^\star}, \mathcal{M}_{\mathbf{\Theta}_t^\star}$. In the third stage, the knowledge of the pretrained teacher system is used to guide and enhance the performance of the student system's encoder and decoder.



The proposed methodology is designed to be general-purpose for CI system design

and serves as an alternative to E2E optimization. In this approach the student is designed to be implementable for real-world acquisition, addressing practical physical constraints in the encoder. However, the teacher system is not required to be practical for real-world acquisition, as it is only used in simulation to transfer its knowledge to the student. This allows for endless possibilities in creating the teacher, such as relaxations in the number of measurements, codification in non-conventional bases (e.g., complex or quaternion), or even using a different CI system as the teacher to distill the student, offering more flexibility than E2E optimization.

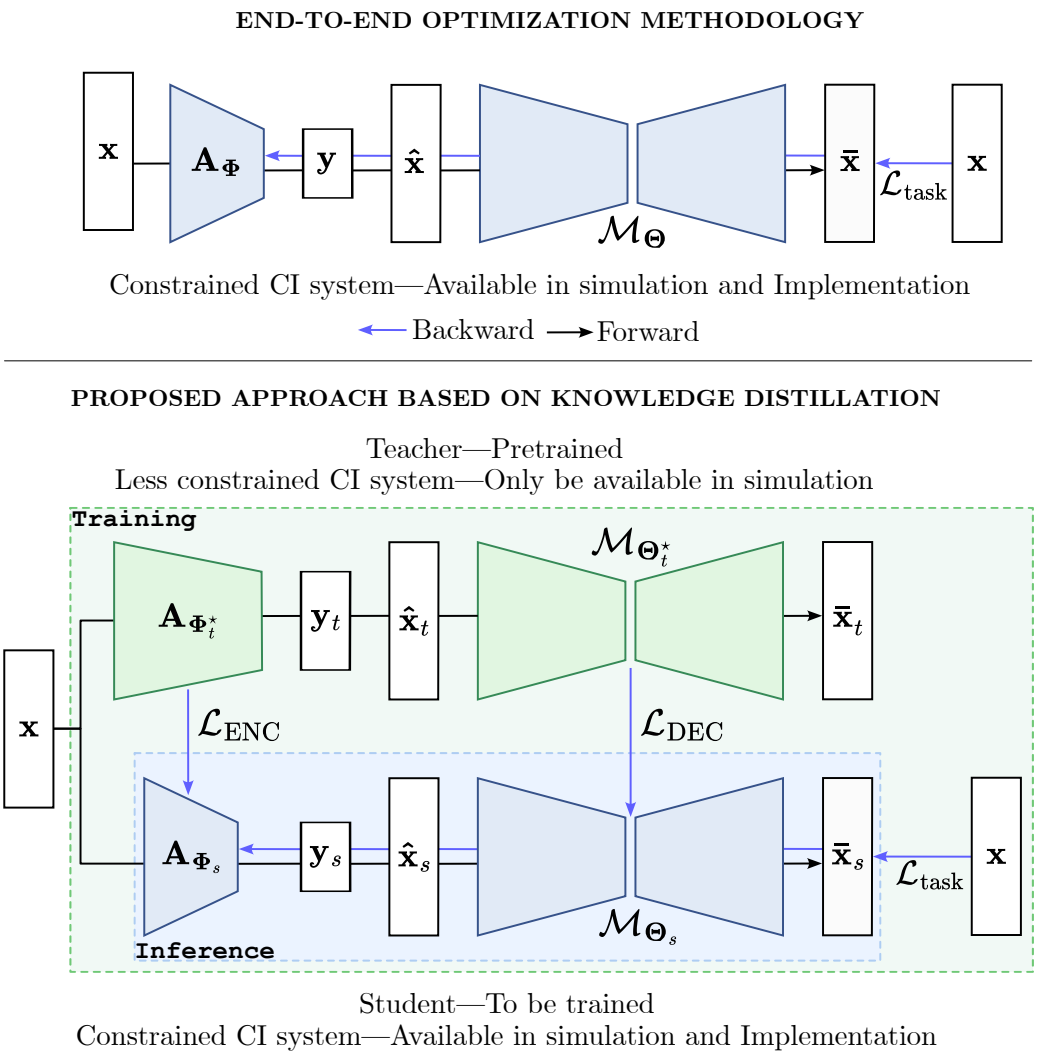
The effectiveness of the proposed approach was validated on three representative CI systems: MRI, the single-pixel camera (SPC) Duarte et al. (2008), and the single-disperser coded aperture snapshot spectral imager (SD-CASSI) Wagadarikar et al. (2008). However, the methodology is not limited to these cases, and future research can further explore its applicability to other imaging modalities such as diffractive optical imaging with DOEs Urrea et al. (2024), computed tomography Liu et al. (2023) and seismic imaging Hernandez-Rojas and Arguello (2024). This research devises different design criteria for relaxing the student's encoder to obtain the teacher's encoder for each system. In MRI, a teacher was created with a lower acceleration factor (AF) than the student, allowing it to acquire more information, which can then be transferred to the student. For the SPC and SD-CASSI systems, the teacher uses real-valued CAs and equal or higher number of snapshots than the binary-valued CA student, enabling the acquisition of more information for transfer. However, the design of the teacher is still an open research question, where different encoder

parametrizations can be useful for knowledge transfer.

To transfer knowledge from the teacher system to the student system, two types of loss functions were employed, as illustrated in Figure 2: the encoder loss function and the decoder loss function. The encoder loss function aligns the student's encoder structure with the teacher's encoder structure, while the decoder loss function aligns the feature space of the student's decoder with the teacher's. Additionally, the encoder loss function not only provides regularization for the student's encoder structure but also enhances the gradient signal that is backpropagated by introducing a supervised signal in the encoder, thereby mitigating the vanishing gradient problem in traditional E2E optimization, similar to the effect of intermediate losses in deep networks Szegedy et al. (2015). The decoder loss guides the learning of optimal intermediate features in the student's computational decoder. This is achieved because the teacher's encoder operates under fewer constraints, enabling higher recovery performance in the teacher's decoder compared to the student's decoder. As a result, the teacher's decoder produces more robust learned feature representations, which can effectively guide the student's feature space. Simulation results indicate that the nature of the teacher system's encoder significantly impacts the performance of the student system. Specifically, a teacher with a relaxation close to that of the student (similar number of measurements or similar nature of the codification basis) provides effective guidance, leading to better encoder design and reconstruction performance than E2E optimization. For MRI, when the undersampling mask learned by the student was used to train a reconstruction neural network, it achieved better reconstruction performance than the mask obtained th-

rough E2E optimization. In the SPC, the student's learned sensing matrix exhibited lower condition numbers and mutual coherence compared to that obtained through E2E optimization. Similarly, in the SD-CASSI, the student's learned encoder led to improved spectral band correlation compared to E2E optimization.

Figure 2. Comparative of E2E optimization and the proposed KD methodology for designing CI systems. During training, the pretrained teacher guides the learning of the student through the proposed loss functions $\mathcal{L}_{\text{ENC}}$ and $\mathcal{L}_{\text{DEC}}$. For inference, the student system operates independently.



2. Problem statement

CI systems involve the joint design of the encoder and computational decoder Bhandari et al. (2022). The state-of-the-art for designing these systems uses an E2E optimization framework, where the encoder is modeled as a neural network layer and is integrated into neural network architectures for performing a specific imaging task, an the parameters of both the encoder and computational decoder are jointly optimized Arguello et al. (2023).

CI systems are constrained by physical practical limitations that must be addressed during its design phase. For instance optical systems use CAs to reduce the amount of acquired information while maximizing reconstruction, furthermore, binary-valued CAs are preferred over real-valued or quantized due to its ease of fabrication, have lower storage requirements and use less integration time. However these constraints limit the performance of the system Bacca et al. (2021). Similarly, in MRI, scans can take long time, leading to discomfort of the patient, motion artifacts and high costs Zaitsev et al. (2015); GharehMohammadi and Sebro (2024), then to address this issues $k-$ space measurements are undersampled, effectively reducing acquisition but compromising image quality Yiasemis et al. (2024).

In order to take into account the physical constraints during the design phase, regularization functions are introduce into the E2E optimization problem, these functions constraint the set of values that the parameters of the encoder can take, effectively addressing practical physical constraints for implementation but reducing de degrees of freedom of the encoder during training, which consequently limits the performance of the system on its task Arguello

et al. (2023). Furthermore, the encoder in the E2E optimization framework is prone to the gradient vanishing gradient problem as it is the first layer of the framework, and intermediate outputs may be suboptimal due to the overall E2E framework is optimized using only the output of the computational decoder Jacome et al. (2023).

In the field of deep learning, the paradigm of KD Hinton et al. (2015) is a popular approach for neural network compression, motivated by compressing high-complexity models into compact networks to enable applications on resource-limited devices Gou et al. (2021). KD is based on a teacher-student relationship, where a large neural network (teacher) is employed to train a smaller neural network (the student) Wang and Yoon (2022), achieving the student higher performance than standalone training. This leads to the following research question:

Can knowledge distillation techniques enhance the optimization of a highly physically constrained computational imaging system, compared to traditional end-to-end deep learning optimization, in terms of performance improvement?

3. Objectives

General Objective

To design a knowledge distillation deep learning algorithm for the optimization of a highly physically constrained coded aperture imaging system.

Specific Objectives

1. To model existing physically constrained coded aperture imaging systems, as well as the knowledge distillation approach as an optimization problem.

2. To develop a knowledge distillation deep learning algorithm for the optimization of a highly physically constrained coded aperture imaging system.

3. To validate the designed algorithm for imaging reconstruction tasks through simulations using different datasets.

4. To compare its performance against traditional E2E deep learning optimization for designing a highly physically constrained coded aperture imaging system.

4. Background and related work

This chapter discusses the state-of-the-art in designing CI systems, namely E2E optimization, provides background on knowledge distillation and its role in CI systems, and describes the CI systems used in this research to validate the proposed method, specifically MRI, the SPC, and the SD-CASSI.

4.1. End-to-end optimization

CI systems can be interpreted as an encoder-decoder framework, where the imaging formation model encodes a scene $\mathbf{x} \in \mathbb{R}^n$ into coded projections $\mathbf{y} \in \mathbb{R}^m$, with $m \ll n$ Arguello et al. (2023). This process follows the forward sensing model:

$$\mathbf{y} = \mathbf{A}_{\mathbf{\Phi}}\mathbf{x} + \boldsymbol{\eta}, \tag{1}$$

where $\mathbf{A}_{\mathbf{\Phi}} \in \mathbb{R}^{m \times n}$ represents the encoder forward model of the CI system, parametrized by $\mathbf{\Phi}$, and $\boldsymbol{\eta} \in \mathbb{R}^m$ is additive noise. A computational decoder then recovers the underlying signal $\mathbf{x}$. The state-of-the-art approach to optimize CI systems involves an E2E deep learning optimization, where the physics-based forward imaging model is incorporated into deep learning architectures as a differentiable layer Arguello et al. (2023); Wetzstein et al. (2020). By modeling both the sensing model $\mathbf{A}_{\mathbf{\Phi}}$ parametrized by $\mathbf{\Phi}$ and the computational decoder $\mathcal{M}_{\mathbf{\Theta}}$ parametrized by $\mathbf{\Theta}$ as neural network layers, the E2E optimization problem is formulated as follows:

$$\{\boldsymbol{\Theta}^{\star}, \boldsymbol{\Phi}^{\star}\} = \underset{\boldsymbol{\Theta},\boldsymbol{\Phi}}{\arg\min}\ \frac{1}{P}\sum_{p=1}^{P} \|\mathcal{M}_{\boldsymbol{\Theta}}(\mathbf{A}_{\boldsymbol{\Phi}}^{\top}\mathbf{A}_{\boldsymbol{\Phi}}\mathbf{x}_p) - \mathbf{x}_p\|_2^2 + \tau\mathcal{R}_{\tau}(\boldsymbol{\Phi}) + \mu\mathcal{R}_{\mu}(\boldsymbol{\Theta}), \tag{2}$$

where $\{\boldsymbol{\Theta}^{\star}, \boldsymbol{\Phi}^{\star}\}$ are the set of optimal values of the encoder and the computational decoder parameters, respectively, $\{\mathbf{x}_p\}_{p=1}^{P}$ is the dataset, and $\mathcal{R}_{\tau}(\boldsymbol{\Phi})$ is a regularization function that acts on the parameters of the imaging formation model to promote specific physical constraints, such as binarization and quantization for CAs in systems like the SPC and SD-CASSI, transmittance for the $k-$space undersampling mask with a given AF in MRI, the number of angles in computed tomography, or the quantization of the height map in DOEs. The regularization parameter τ balances the trade-off between the desired constraints promoted by the regularization function and the performance of the recovery Arguello et al. (2023). Additionally, the regularization function $\mathcal{R}_{\mu}(\boldsymbol{\Theta})$ prevents overfitting of the computational decoder parameters $\boldsymbol{\Theta}$ to the training data by limiting the network's complexity through penalizing large weight values, where μ as its regularization parameter.

The regularization functions $\mathcal{R}_{\tau}(\boldsymbol{\Phi})$ used to enforce the physical constraints of the CI systems in the E2E optimization, restrict the set of optimal values in the imaging formation model. Consequently, this reduces the degrees of freedom during training, ultimately limiting the recovery performance of the CI system. This performance reduction occurs due to the influence of physical constraints on the structure of the imaging formation model, which affects the effectiveness of the computational decoder Bacca et al. (2021); Jacome et al.

(2023). Furthermore, the encoder is susceptible to vanishing gradients as it is the first layer in the E2E framework, which further limits its ability to learn an optimal codification. Therefore, effective design of the encoder is essential for achieving high-performance recovery in CI systems.

Equation (2) is solved using gradient descent-based algorithms, and once the optimal encoder values are learned $\mathbf{\Phi}^{\star}$, the physical imaging formation model can be implemented to acquire real measurements. Subsequently, the learned computational decoder $\mathcal{M}_{\mathbf{\Theta}^{\star}}$ is used to perform the recovery task on the acquired measurements Arguello et al. (2023); Wetzstein et al. (2020).

4.2. Knowledge distillation

Deploying large-scale deep learning models poses a significant challenge due to their immense computational complexity and massive storage requirements, limiting their use in resource-constrained systems Cho and Hariharan (2019); Gou et al. (2021). To achieve similar performance while reducing model complexity, a model compression technique is needed.

KD, popularized by Hinton et al. (2015), is an effective technique for optimizing neural networks for inference on resource-constrained devices. In KD a smaller and simpler model, known as the student, is trained to mimic the behavior of a larger and more complex model, known as the teacher, leading to the student model achieving better performance than would be possible if it were trained directly on the task Gou et al. (2021); Romero et al. (2015). KD is especially valuable in scenarios where the deployment of neural network models for real-time inference is challenging due to physical constraints or system limitations, such as

in resource-constrained devices like smartphones, smartwatches, or embedded devices Mait et al. (2018), where storage and computational resources are limited.

Knowledge transfer in KD can occur through various sources, including intermediate feature maps of the teacher model Romero et al. (2015), relationships between consecutive feature maps Yim et al. (2017), and the response of the teacher model Hinton et al. (2015). Feature maps represent outputs from intermediate layers, while the response refers to the output of the model's final layer Alkhulaifi et al. (2020). The KD optimization problem can be expressed as:

$$\boldsymbol{\Theta}_s^\star = \arg\min_{\boldsymbol{\Theta}_s} \frac{1}{P} \sum_{p=1}^{P} \lambda_1 \mathcal{L}_{\text{task}}(\mathcal{M}_{\boldsymbol{\Theta}_s}(\mathbf{x}_p), \mathbf{d}_p) + \lambda_2 \mathcal{L}_{\text{KD}}(\boldsymbol{\Theta}_s, \boldsymbol{\Theta}_t^\star, \mathbf{x}_p) \tag{3}$$

where $\boldsymbol{\Theta}_s$ represents the parameters of the student model, $\boldsymbol{\Theta}_t^\star$, denotes the optimized parameters of the teacher model, $\mathbf{x}_p$ and $\mathbf{d}_p$ are the input and desired output of the model, $\mathcal{L}_{\text{task}}$ is the task-specific loss (e.g., reconstruction or classification), and $\mathcal{L}_{\text{KD}}$ is the KD loss that transfers knowledge from the teacher to the student. The regularization parameters λ_1 and λ_2 balance the contributions of the task-specific and KD losses.

Although KD has been applied mainly to high-level computer vision tasks such as segmentation, detection, and classification, its applications in low-level vision tasks such as reconstruction and denoising have been less explored Fang et al. (2023); Zhu et al. (2023); Gao et al. (2019); He et al. (2020); Xie et al. (2023); Chen et al. (2021). Furthermore, the use of KD in CI has received limited attention. For example, Wu and Li (2024) applied KD to

reconstruct magnetic resonance images but did not address the design of the undersampling mask. Similarly, Quan et al. (2023) utilized KD to enhance the performance of a phase retrieval system's recovery network without considering the design of the codification masks, while Sun et al. (2024) employed KD for compressive video captioning without optimizing the CAs.

4.3. Acquisition models

This section describes the CI systems used to validate the proposed approach. These systems include three popular CI systems: MRI, the SPC, and the SD-CASSI. The following subsections provide detailed descriptions of each system.

4.3.1. Magnetic resonance imaging. MRI is a non-invasive imaging technique that captures images of the body using strong magnetic fields and radio waves. Unlike X-ray imaging, it does not employ harmful ionizing radiation Karl et al. (2023). MRI scans can take a long time depending on the part of the body being imaged, which can cause discomfort to the patient, increase costs, and cause motion artifacts. Additionally, this extended scan time limits the number of patients that can be imaged in a day, resulting in long waiting times that can stretch for weeks or even months Bahadir et al. (2020); GharehMohammadi and Sebro (2024). Therefore, the acquisition process is undersampled to reduce scan times, and magnetic resonance images are reconstructed from the undersampled $k-$space measurements. Traditional undersampling approaches use fixed undersampling patterns, with the most common including random rectilinear, equispaced rectilinear, cartesian, radial, and spiral Yiasemis et al. (2024). However, each undersampling pattern may lead

to different reconstruction performances for the same reconstruction method Bahadir et al. (2020); Geethanath et al. (2013); Yiasemis et al. (2024). More recent approaches optimize the undersampling masks pattern together with the reconstruction neural network in an E2E manner Bahadir et al. (2020); Weber et al. (2024); Weiss et al. (2020); Razumov et al. (2023).

The single-coil MRI acquisition process is given by equation (1), where $\mathbf{A}_{\mathbf{\Phi}} = \mathbf{\Phi}\mathbf{F}$, where $\mathbf{\Phi} \in \{0,1\}^{m\times n}$ denotes an undersampling mask and $\mathbf{F} \in \mathbb{C}^{n\times n}$ is the Fourier transform operator. In this work, the $k-$space undersampling mask is modeled as a neural network layer with trainable parameters $\mathbf{W} \in \mathbb{R}^{m\times n}$, incorporating a Heaviside step activation function $H(x) = \frac{x+|x|}{2|x|}$ to binarize the mask, such that $\mathbf{\Phi} = H(\mathbf{W}) \in \{0,1\}^{m\times n}$. However, this activation function is not differentiable at zero and has a gradient of zero elsewhere, making it unsuitable for learning the undersampling mask directly. To address this, a Straight-Through Estimator (STE) is employed Courbariaux et al. (2015). The STE enables the gradient from the preceding layer to propagate through the activation function, facilitating the optimization of the binary undersampling mask, such that $\frac{\partial \mathbf{\Phi}}{\partial \mathbf{W}} = \mathbf{I}$.

To promote the desired acceleration factor for the $k-$space undersampling mask $\mathbf{\Phi}$ the following regularization term is employed in (2) as:

$$\mathcal{R}_\tau(\mathbf{\Phi}) = \tau \left(\frac{\|\mathbf{\Phi}\|_0}{n} - \frac{1}{AF} \right)^4. \tag{4}$$

4.3.2. Single pixel camera. The SPC consists of an objective lens that forms an image of the scene onto a CA, which spatially modulates the scene. The encoded

Figure 3. Single-pixel camera acquisition scheme

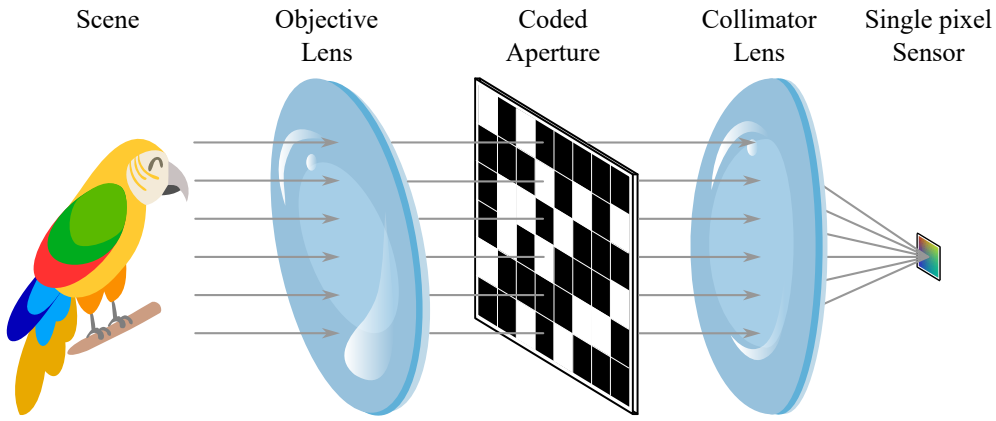
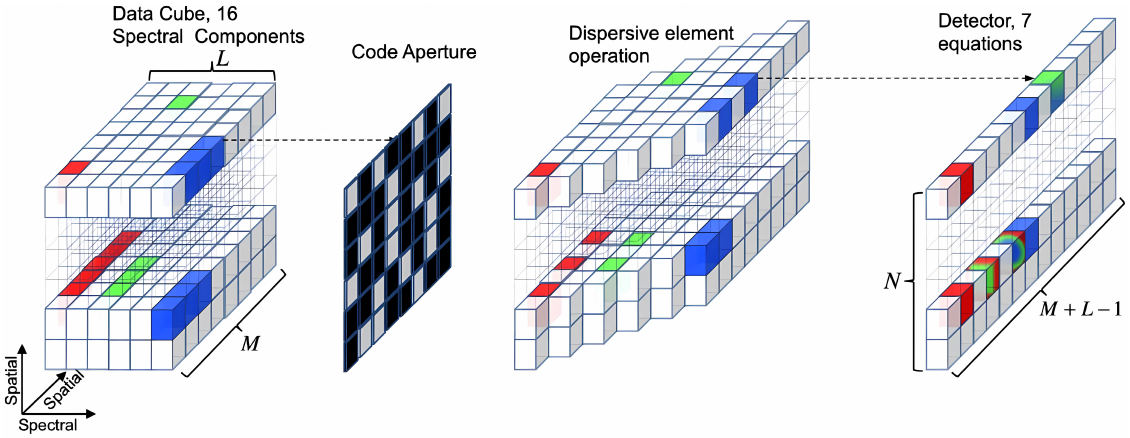


image is then focused by a collimator lens onto a single-photon detector, as illustrated Figure 3. To obtain different measurements (snapshots) from the same scene, the CA is changed for each snapshot Bacca et al. (2019).

The forward model of the SPC is given by equation (1). Here, the forward operator matrix is built upon the vectorization of the CAs $\mathbf{\Phi}_i$ as $\mathbf{A}_{\mathbf{\Phi}} = [\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_m]^\top$. The relationship between the number of acquired measurements and the image size is given by the compression ratio $\gamma = \frac{m}{n}$.

The CAs can be binary-valued or real-valued. In a binary-valued CA, translucent or opaque elements are used to either block or allow light to pass through. Each entry of the CA takes a value of 0 or 1 ($\mathbf{\Phi}_{i,j} \in \{0, 1\}$), representing the blocking or passage of light Bacca et al. (2020); Arce et al. (2014). On the other hand, real-valued CAs attenuate light at different levels, admitting a wider range of values compared to binary CAs. In a real-valued CA, each entry takes a value $\mathbf{\Phi}_{i,j} \in \mathbb{R}$ Bacca et al. (2021). The SPC system aims to minimize the number of snapshots, reducing acquisition and processing times while

Figure 4. Single-disperser coded aperture snapshot spectral imager acquisition scheme, source Henry and Gonzalo (2011).



maximizing reconstruction performance.

In this work, the CAs of the SPC were modeled as neural network layers with trainable parameters $\mathbf{W}_i \in \mathbb{R}^n, i = 1, \ldots, m$. The binary-valued CAs are obtained as $\mathbf{\Phi}_i = \text{sign}(\mathbf{W}_i) \in \{-1, 1\}^n, i = 1, \ldots, m$ where $\text{sign}(\cdot)$ denotes the element-wise sign function $\text{sign}(x) = \frac{x}{|x|}$. However, this function, like the Heaviside step, is not differentiable at zero and has a gradient of zero elsewhere, leading to the same optimization issues. Therefore, we use an STE, such that $\frac{\partial \mathbf{\Phi}}{\partial \mathbf{W}} = \mathbf{I}$.

4.3.3. Single disperser coded aperture snapshot spectral imager. The SD-CASSI system Wagadarikar et al. (2008) is a snapshot spectral imaging system where the incoming light is modulated spatially using a CA. Subsequently, a disperser element, such as a prism or a grating, disperses the spectral information of each spatial location across a wide area on the detector, resulting in the multiplexation of the spatial and spectral information of the scene, as illustrated in Figure 4.

The forward model of the SD-CASSI is expressed by equation (1). Here $\mathbf{x} \in \mathbb{R}^{NML}$ represents the vectorized data cube with spatial dimensions $N \times M$ and L spectral bands with $n = NML$, $\mathbf{y} \in \mathbb{R}^{N(M+L-1)}$ denotes the vectorized detector measurement with $m = N(M + L-1)$, and $\boldsymbol{\eta} \in \mathbb{R}^{N(M+L-1)}$ is additive noise. The sensing matrix that accounts for the coding and dispersing effect is given by $\mathbf{A}_{\mathbf{\Phi}} = \mathbf{P}\mathbf{\Phi} \in \mathbb{R}^{N(M+L-1)\times NML}$, where $\mathbf{P} \in \mathbb{R}^{N(M+L-1)\times NML}$ is a matrix that models the dispersion and $\mathbf{\Phi} \in \mathbb{R}^{NML\times NML}$ is the diagonalized CA. To acquire a set of k snapshots $\mathbf{y} = [(\mathbf{y}^0)^\top, \ldots, (\mathbf{y}^{k-1})^\top]^\top$ represents the set of measurements, $\mathbf{A}_{\mathbf{\Phi}} = [(\mathbf{A}^0_{\mathbf{\Phi}})^\top, \ldots, (\mathbf{A}^{k-1}_{\mathbf{\Phi}})^\top]^\top$ are the set of sensing matrices, and $\boldsymbol{\eta} = [(\boldsymbol{\eta}^0)^\top, \ldots, (\boldsymbol{\eta}^{k-1})^\top]^\top$ is the set of noise vectors. The SD-CASSI shares similar constraints with the SPC system. In the SD-CASSI, CAs may be binary-valued ($\mathbf{\Phi}_{i,j} \in \{0, 1\}$) or real-valued ($\mathbf{\Phi}_{i,j} \in [0, 1]$), additionally, the number of snapshots is minimized to reduce acquisition and processing times while maximizing recovery performance. To model the CAs for the SD-CASSI in the E2E framework, the same process done for the SPC is applied using the Heavyside step activation function H along with a STE.

5. Distilling knowledge for computational imaging system design

This chapter introduces the proposed approach to address the performance limitations of CI systems optimized through E2E optimization. Inspired by KD, this approach enables knowledge transfer from a less constrained CI system (teacher) to guide the learning of a more constrained CI system (student). Specifically, the proposed approach is divided into three stages. First, the teacher system encoder is obtained as a relaxation of the student system encoder, such that $\mathbf{A}_{\mathbf{\Phi}_t} \leftarrow \mathcal{Z}(\mathbf{A}_{\mathbf{\Phi}_s})$, where $\mathcal{Z}(\cdot)$ is a relaxation operator. Second, the teacher is optimized following Eq. (2), solving a relaxed version of the student's recovery problem, obtaining the optimal parameters of the encoder and decoder $\{\mathbf{\Theta}_t^\star, \mathbf{\Phi}_t^\star\}$. Third, once trained, the teacher guides the student's learning process through proposed knowledge transfer functions. The proposed KD optimization pipeline is illustrated in algorithm 1.

Before describing the proposed KD optimization scheme, the nature of the teacher encoder is first defined, and its key differences from the student encoder are outlined. Next, we describe the employed computational decoder.

5.1. Computational encoder

The design of the encoder of a CI system is the most important part of the system, as it determines how the scene is sampled and encoded. This directly affects the amount and quality of the acquired information, which is critical for the performance of the computational decoder. We now describe the teacher and student encoder configurations.

Algorithm 1 Proposed KD optimization of CI systems

Require: $\mathbf{A}_{\mathbf{\Phi}_s}, \mathcal{M}_{\mathbf{\Theta}_s}, N$ ▷ Student's encoder and computational decoder, number of epochs

Stage 1: Obtain the teacher's encoder $\mathbf{A}_{\mathbf{\Phi}_t}$ by relaxing the student's encoder constraints with operator $\mathcal{Z}(\cdot)$:

$\mathbf{A}_{\mathbf{\Phi}_t} \leftarrow \mathcal{Z}(\mathbf{A}_{\mathbf{\Phi}_s})$

Stage 2: Optimize the teacher's encoder and decoder parameters

for i in $1, 2, \ldots, N$ **do**

$\mathcal{L}_{\text{task}} = \left\| \mathcal{M}_{\mathbf{\Theta}_t^i} \left(\mathbf{A}_{\mathbf{\Phi}_t^i}^{\top} \mathbf{A}_{\mathbf{\Phi}_t^i} \mathbf{x} \right) - \mathbf{x} \right\|_2^2$

$\mathcal{L} = \mathcal{L}_{\text{task}} + \tau \mathcal{R}_\tau(\mathbf{\Phi}_t^i) + \mu \mathcal{R}_\mu(\mathbf{\Theta}_t^i)$

$\mathbf{\Theta}_t^{i+1} = \mathbf{\Theta}_t^i - \alpha \left(\frac{\partial \mathcal{L}_{\text{task}}}{\partial \mathbf{\Theta}_t^i} + \mu \frac{\partial \mathcal{R}_\mu(\mathbf{\Theta}_t^i)}{\partial \mathbf{\Theta}_t^i} \right)$

$\mathbf{\Phi}_t^{i+1} = \mathbf{\Phi}_t^i - \alpha \left(\frac{\partial \mathcal{L}_{\text{task}}}{\partial \mathbf{\Phi}_t^i} + \tau \frac{\partial \mathcal{R}_\tau(\mathbf{\Phi}_t^i)}{\partial \mathbf{\Phi}_t^i} \right)$

end for

The teacher's encoder and decoder parameters $\{\mathbf{\Theta}_t^\star, \mathbf{\Phi}_t^\star\}$ are set to be frozen.

Stage 3: Optimize $\mathbf{A}_{\mathbf{\Phi}_s}, \mathcal{M}_{\mathbf{\Theta}_s}$ with the guidance of $\mathbf{A}_{\mathbf{\Phi}_t^\star}, \mathcal{M}_{\mathbf{\Theta}_t^\star}$ using knowledge transfer functions $\mathcal{L}_{\text{ENC}}$ and $\mathcal{L}_{\text{DEC}}$.

for i in $1, 2, \ldots, N$ **do**

$\mathbf{S} = \mathcal{M}_{\mathbf{\Theta}_s}^{[l]}(\mathbf{A}_{\mathbf{\Phi}_s^i}^{\top} \mathbf{A}_{\mathbf{\Phi}_s^i} \mathbf{x})$ ▷ Student's features at layer l

$\mathbf{T} = \mathcal{M}_{\mathbf{\Theta}_t^\star}^{[l]}(\mathbf{A}_{\mathbf{\Phi}_t^{i\star}}^{\top} \mathbf{A}_{\mathbf{\Phi}_t^{i\star}} \mathbf{x})$ ▷ Teacher's features at layer l

$\mathcal{L}_{\text{ENC}} = \mathcal{D}\left(\mathbf{A}_{\mathbf{\Phi}_s^i}, \mathbf{A}_{\mathbf{\Phi}_t^{i\star}}, \mathbf{x} \right)$ ▷ $\mathcal{D}$ is a distance function

$\mathcal{L}_{\text{DEC}} = \|\mathbf{S} - \mathbf{T}\|_2^2$

$\mathcal{L}_{\text{task}} = \left\| \mathcal{M}_{\mathbf{\Theta}_s^i} \left(\mathbf{A}_{\mathbf{\Phi}_s^i}^{\top} \mathbf{A}_{\mathbf{\Phi}_s^i} \mathbf{x} \right) - \mathbf{x} \right\|_2^2$

$\mathcal{L}_{\text{KD}} = \lambda_1 \mathcal{L}_{\text{task}} + \lambda_2 \mathcal{L}_{\text{DEC}} + \lambda_3 \mathcal{L}_{\text{ENC}} + \tau \mathcal{R}_\tau(\mathbf{\Phi}_s^i) + \mu \mathcal{R}_\mu(\mathbf{\Theta}_s^i)$

$\mathbf{\Theta}_s^{i+1} = \mathbf{\Theta}_s^i - \alpha \left(\lambda_1 \frac{\partial \mathcal{L}_{\text{task}}}{\partial \mathbf{\Theta}_s^i} + \lambda_2 \frac{\partial \mathcal{L}_{\text{DEC}}}{\partial \mathbf{\Theta}_s^i} + \mu \frac{\partial \mathcal{R}_\mu(\mathbf{\Theta}_s^i)}{\partial \mathbf{\Theta}_s^i} \right)$

$\mathbf{\Phi}_s^{i+1} = \mathbf{\Phi}_s^i - \alpha \left(\lambda_1 \frac{\partial \mathcal{L}_{\text{task}}}{\partial \mathbf{\Phi}_s^i} + \lambda_2 \frac{\partial \mathcal{L}_{\text{DEC}}}{\partial \mathbf{\Phi}_s^i} \right) + \alpha \left(\lambda_3 \frac{\partial \mathcal{L}_{\text{ENC}}}{\partial \mathbf{\Phi}_s^i} + \tau \frac{\partial \mathcal{R}_\tau(\mathbf{\Phi}_s^i)}{\partial \mathbf{\Phi}_s^i} \right)$

end for

return $\{\mathbf{A}_{\mathbf{\Phi}_s^\star}, \mathcal{M}_{\mathbf{\Theta}_s^\star}\}$

5.1.1. Teacher encoder nature. A key aspect of the proposed approach is the teacher configuration. The teacher's encoder is obtained by relaxing the student's encoder constraints using the operator $\mathcal{Z}(\cdot)$, such that $\mathbf{A}_{\mathbf{\Phi}_t} \leftarrow \mathcal{Z}(\mathbf{A}_{\mathbf{\Phi}_s})$. Furthermore, the teacher is a synthetic CI system used only in simulations to guide the student, and therefore, it does

not need to be feasible for real-world acquisition. This flexibility enables the exploration of various teacher configurations based on the application. As a result, the teacher's encoder codifies richer information than the student's, enabling its computational decoder to achieve higher performance. Below, we mention the explored teacher encoder configurations:

- **MRI:** The teacher system is obtained by relaxing the student's acceleration factor, i.e, $\mathbf{\Phi}_t \in \{0,1\}^{m_t \times n}, \mathbf{\Phi}_s \in \{0,1\}^{m_s \times n}$, with $\frac{n}{m_s} > \frac{n}{m_t}$.
- **SPC:** The teacher system is obtained by relaxing the student's constraint on the CAs from binary-valued to real-valued ($[\mathbf{\Phi}_t]_{i,j} \in \mathbb{R}$), and increasing the amount of acquired information, such that $m_t \geq m_s$.
- **SD-CASSI:** The teacher systems follow the same relaxations explored for the SPC.

Remark 1 *Here we devised a set of design criteria for the teacher model in the employed CI systems, however, this is still an open research question. We devise some insights that can be valuable for other imaging modalities. For computed tomography, where we are required to design the source angles to be minimal to reduce the user's radiation exposure Mao et al. (2018) or in scenarios of limited angle where only a small portion of the scene can be scanned Liu et al. (2023), a teacher can employ a huge number of equispaced source angles or a higher resolution sensor. In diffractive optical imaging, the DOEs have quantized heights Li et al. (2022) or in some scenarios DOEs are implemented in deformable mirrors which have a very limited parametrization based on few Zernike polynomials Urrea et al. (2024) which reduces the degree of freedom in the optimization. In this scenario, the teacher could be*

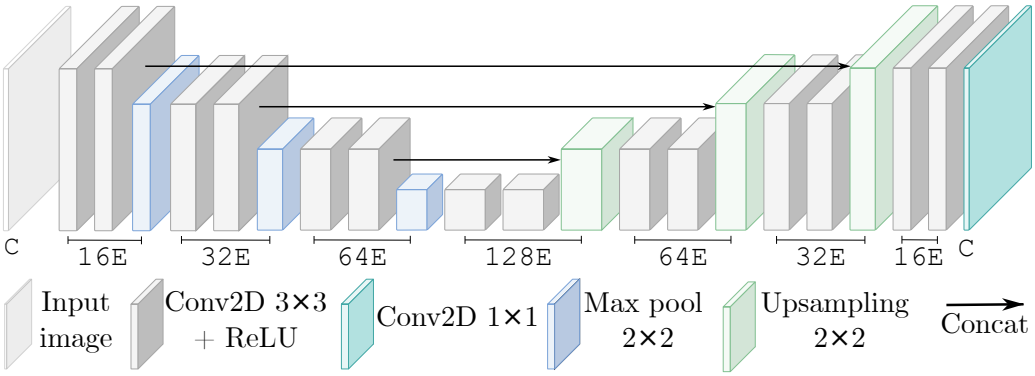
a quantization-free DOE or employ a high number of Zernike polynomials. In compressive seismic imaging, one is required to reduce the number of sources to mitigate environmental effects and acquisition costs and time Hernandez-Rojas and Arguello (2024). Thus, a teacher design can be a system that uses a high number of uniformly distributed sources.

5.1.2. Student encoder nature. The student CI system encoder, designed for both simulation and real-world implementation, must account for practical constraints. In MRI, we considered students with undersampled $k-$space masks corresponding to acceleration factors $AF_s \in \{4, 8, 16\}$. For SPC, the following compression ratios were considered, with the student using binary-valued CAs: $\gamma_s \in \{0.05, 0.1, 0.2, 0.3, 0.4\}$. For the SD-CASSI system, the focus is on a student with a single snapshot and a binary-valued CA.

5.2. Computational decoder

Since we focus on developing a new methodology for designing CI systems rather than creating a novel signal reconstruction architecture, we used the well-known U-Net Ronneberger et al. (2015) as the computational decoder for the experimental simulations. However, we emphasize that the proposed methodology can be extended to any computational decoder. Additionally, as our focus is on addressing the physical constraints in the encoder of the student, the computational decoders of both the student and teacher share the same U-Net architecture. The U-Net architecture used in these simulations is depicted in Figure 5.

Figure 5. Scheme of the U-Net used as computational decoder, `C` is the number of channels of the input image, `C=2` for MR images, `C=1` for grayscale images, and `C=8` for the multi-spectral images. `E` determines the number of filters of each convolutional layer of the U-Net, `E=1` for MRI, `E=4` for the SPC, and `E=2` for the SD-CASSI system.



5.3. Knowledge transfer functions

The transfer of knowledge from the teacher to the student is illustrated in Figure 2. In this approach, the pretrained teacher system guides the student's learning through two proposed types of loss functions: the encoder loss function $\mathcal{L}_{\text{ENC}}$ and the decoder loss function $\mathcal{L}_{\text{DEC}}$. Specifically, $\mathcal{L}_{\text{DEC}}$ promotes the alignment of the intermediate feature spaces between the teacher and student decoders, while $\mathcal{L}_{\text{ENC}}$ enforces alignment between the teacher and student encoders. These two functions allow guidance in both the sensing architecture and the computational recovery method, mitigating the vanishing gradient problem in the encoder and facilitating the learning of optimal intermediate outputs in the computational decoder. The learning of the student system involves minimizing the following loss function:

$$\mathcal{L}_{\text{KD}} = \lambda_1 \|\mathcal{M}_{\boldsymbol{\Theta}_s}(\mathbf{A}_{\boldsymbol{\Phi}_s}^{\top}\mathbf{A}_{\boldsymbol{\Phi}_s}\mathbf{x}) - \mathbf{x}\|_2^2 + \lambda_2\mathcal{L}_{\text{DEC}} + \lambda_3\mathcal{L}_{\text{ENC}}, \tag{5}$$

where $\lambda_3 = 1 - \lambda_1 - \lambda_2$ with $\lambda_3 > 0$. The parameters λ_1, λ_2, and λ_3 are regularization terms that control the relative contributions of the different loss functions during optimization. Then, the KD optimization problem is then formulated as follows:

$$\begin{aligned}\{\boldsymbol{\Theta}_s^\star, \boldsymbol{\Phi}_s^\star\} =& \underset{\boldsymbol{\Theta}_s, \boldsymbol{\Phi}_s}{\arg\min} \frac{1}{P} \sum_{p=1}^{P} \lambda_1 \|\mathcal{M}_{\boldsymbol{\Theta}_s}(\mathbf{A}_{\boldsymbol{\Phi}_s}^\top \mathbf{A}_{\boldsymbol{\Phi}_s} \mathbf{x}_p) - \mathbf{x}_p\|_2^2 \\ &+ \lambda_2 \mathcal{L}_{\text{DEC}} + \lambda_3 \mathcal{L}_{\text{ENC}} + \tau \mathcal{R}_\tau(\boldsymbol{\Phi}_s) + \mu \mathcal{R}_\mu(\boldsymbol{\Theta}_s).\end{aligned} \tag{6}$$

The decoder KD loss function, $\mathcal{L}_{\text{DEC}}$, remains consistent across all three CI systems. This loss focuses on aligning the feature space of the student and teacher computational decoders. Since the teacher network contains clearer and more robust features—having been trained to reconstruct images with fewer degradations due to having a less constrained encoder than the student—its feature space can guide and improve the student's computational decoder performance. To this end, we use the bottleneck layers, as they provide the most compact and informative representation of the input image. The decoder loss function is defined as:

$$\mathcal{L}_{\text{DEC}} = \frac{1}{B} \|\mathbf{T}_b - \mathbf{S}_b\|_2^2, \tag{7}$$

where $\mathbf{T}_b = \mathcal{M}_{\boldsymbol{\Theta}_t^\star}^{[b]}(\mathbf{A}_{\boldsymbol{\Phi}_t^\star}^\top \mathbf{A}_{\boldsymbol{\Phi}_t^\star} \mathbf{x})$ and $\mathbf{S}_b = \mathcal{M}_{\boldsymbol{\Theta}_s}^{[b]}(\mathbf{A}_{\boldsymbol{\Phi}_s}^\top \mathbf{A}_{\boldsymbol{\Phi}_s} \mathbf{x})$ represent the intermediate feature maps extracted from the bottleneck layers ($[b]$) of the teacher and student decoders, respectively.

Since the encoder depends on the structure of each acquisition system (MRI, SPC,

SD-CASSI), the encoder loss is defined for each CI system:

5.3.1. Magnetic resonance imaging. The encoder loss function proposed for MRI systems is:

$$\mathcal{L}_{\text{ENC}} = \frac{1}{B} \left\| \mathbf{F}^{\text{H}} \mathbf{\Phi}_s^{\top} \mathbf{\Phi}_s \mathbf{F}\mathbf{x} - \mathbf{F}^{\text{H}} \mathbf{\Phi}_t^{\star\top} \mathbf{\Phi}_t^{\star} \mathbf{F}\mathbf{x} \right\|_2^2, \tag{8}$$

where $\mathbf{F}^{\text{H}}$ is the transpose conjugate of the Fourier transform matrix $\mathbf{F}$. This function encourages the student's $k-$space undersampling mask $\mathbf{\Phi}_s$ to mimic the structure of the teacher's $k-$space undersampling mask $\mathbf{\Phi}_t^{\star}$ by aligning the backprojected measurements from the student's and teacher's encoders in the image domain.

5.3.2. Single-pixel camera. For the SPC system, the encoder loss function is defined as:

$$\mathcal{L}_{\text{ENC}} = \frac{1}{B} \left\| \mathbf{A}_{\mathbf{W}_s}^{\top} \mathbf{A}_{\mathbf{W}_s} - \mathbf{A}_{\mathbf{\Phi}_t^{\star}}^{\top} \mathbf{A}_{\mathbf{\Phi}_t^{\star}} \right\|_2^2, \tag{9}$$

where $\mathbf{A}_{\mathbf{W}s}$ is the student's forward model with CAs before the binarization step, and $\mathbf{A}_{\mathbf{\Phi}_t^{\star}} \in \mathbb{R}^{m \times n}$ is the teacher's forward model, with the teacher employing real-valued CAs, $\mathbf{\Phi}_t^{\star} = \mathbf{W}_t^{\star}$. This loss encourages structural similarity in the student's forward model with the teacher's by aligning their Gram matrices, which capture the pairwise correlations of the CA patterns. The alignment is performed in the real-valued space, where the student's CAs are comparable to the teacher's CAs. By comparing the models in the real-valued domain before binarization, the student can more effectively mimic the teacher's encoding, leveraging

the richer information present in real numbers to guide the learning process.

5.3.3. Single disperser coded aperture snapshot spectral imager. For the SD-CASSI system, the encoder loss function is expressed as:

$$\mathcal{L}_{\text{ENC}} = \frac{1}{B}\|\mathbf{W}_s^\top \mathbf{W}_s - \mathbf{\Phi}_t^{\star\top}\mathbf{\Phi}_t^\star\|_2^2, \tag{10}$$

where $\mathbf{W}_s$ is the student CA prior binarization, and $\mathbf{\Phi}_t^\star$ is the optimal teacher real-valued CA. This function promotes structural similarity between the Gram matrices of the student and teacher CAs. Similar to the SPC system, this approach aligns the student's CAs in the real-valued space, where they are directly comparable to the teacher's CAs.

6. Simulations & Results

Reconstruction performance was measured using the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) Wang et al. (2004). The teacher and baseline models for all three CI systems were previously optimized with E2E optimization according to equation (2). All students were trained under the guidance of their optimal teacher, corresponding to a less constrained CI system with a similar number of snapshots and a comparable nature of the codification basis of the encoder. For specific details on how the optimal teachers were selected, refer to the Section 7.3 of the Ablation Studies Chapter.

Furthermore, the improvement in the encoder design was evaluated for each CI system. In MRI we utilized the learned undersampling mask with the baseline and student as fixed undersampling patterns to train a reconstruction network. We computed the condition number, singular value distribution, and gram matrices of the forward models for the SPC. In the SD-CASSI system, we computed the spectral band correlation of the optimized forward model and the magnitude of the Fourier transform of the designed CA. Additional experiments, including computational requirements, loss function ablation studies, teacher system selection, and noise robustness evaluations, are detailed in the Ablation Studies Chapter. The implementation employs PyTorch Paszke et al. (2019) for deep learning components and Colibri [1] for optical system simulations.

---

[1] **https://github.com/pycolibri/pycolibri**

The following subsections describe the simulations and results for each CI system employed.

6.1. Magnetic resonance imaging

Training details: The FastMRI single-coil knee dataset Knoll et al. (2020), obtained from the DeepInverse library Tachella et al. (2023), was used. It consists of magnetic resonance images of knees, each with a resolution of 320×320. The dataset includes 900 training images, of which 810 were used for training and 90 for validation. The test set contains 73 images. All images were resized to 256×256. The teacher, student, and baseline models were trained for 500 epochs. The first 200 epochs constituted an unconstrained phase, where the acceleration factor regularization function in (4) used a regularization parameter $\tau = 1$, allowing the exploration of different undersampling schemes. The remaining 300 epochs formed a constraining and refining phase, with the regularization parameter set to $\tau = 1 \times 10^{15}$ to limit the mask $\mathbf{\Phi}$ to the given acceleration factor AF and refine the decoder network training with the learned mask. The training was conducted with a 5×10^{-4} learning rate and a batch size of $B = 32$. The AdamW Loshchilov (2017) optimizer was employed with a weight decay parameter of 1×10^{-2}. For the following results, we set $\lambda_1 = 0.1, \lambda_2 = 0.3$ for $AF_s = 4$, $\lambda_1 = 0.3, \lambda_2 = 0.2$ for $AF_s = 8$, and $\lambda_1 = 0.3, \lambda_2 = 0.5$ for $AF_s = 16$. These values were chosen based on a hyperparameter search detailed in Section 7.1 of the Ablation Studies Chapter.

We compared the proposed KD approach with the traditional E2E optimization scheme for designing MRI systems. The teacher systems were previously optimized with an

acceleration factor $AF_t = AF_s - 1$ (for details of this selection, refer to Section 7.3 of the Ablation Studies Chapter). Five realizations of the student MRI systems and baselines were performed, with the average and standard deviation of these realizations reported. Table 1 summarizes the obtained results. It can be observed that in all three evaluated acceleration factors, the student achieves better results than the baseline. Furthermore, the results indicate that all student models are more stable than the baseline, exhibiting lower standard deviations in the reconstruction metrics.

Figure 6 shows visual results comparing the proposed method with the E2E baseline in terms of reconstruction performance and undersampling $k-$space mask design. As the figure illustrates, the student's undersampling mask effectively imitates the teacher's undersampling mask across different acceleration factors, due to the use of the $\mathcal{L}_{\text{ENC}}$ function, yielding better reconstructions than the baseline, getting improvement by up to 1.47 (dB) in PSNR. Additionally, zoom-in sections demonstrate the student model's superior ability to recover high-frequency details and preserve fine structural information than the baseline.

Figure 6. Reconstruction performance of the student and baseline MRI systems. The first row shows the teacher-optimized undersampling mask and its corresponding reconstruction. The second row presents the student-optimized mask and its reconstruction. The third row displays the baseline-optimized mask and its reconstruction. The PSNR (dB) and SSIM metrics are reported in the upper-right corner of each reconstruction.

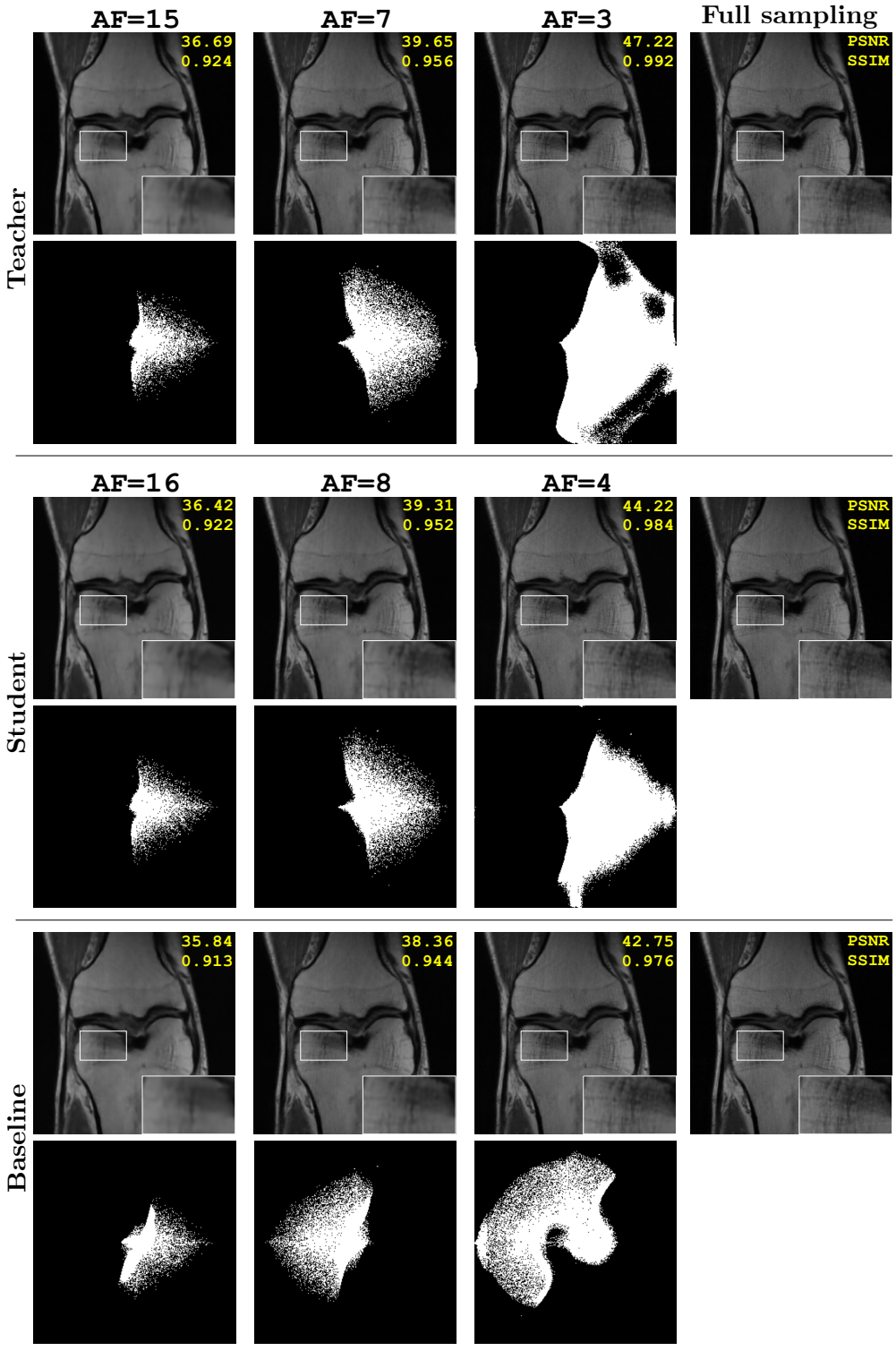
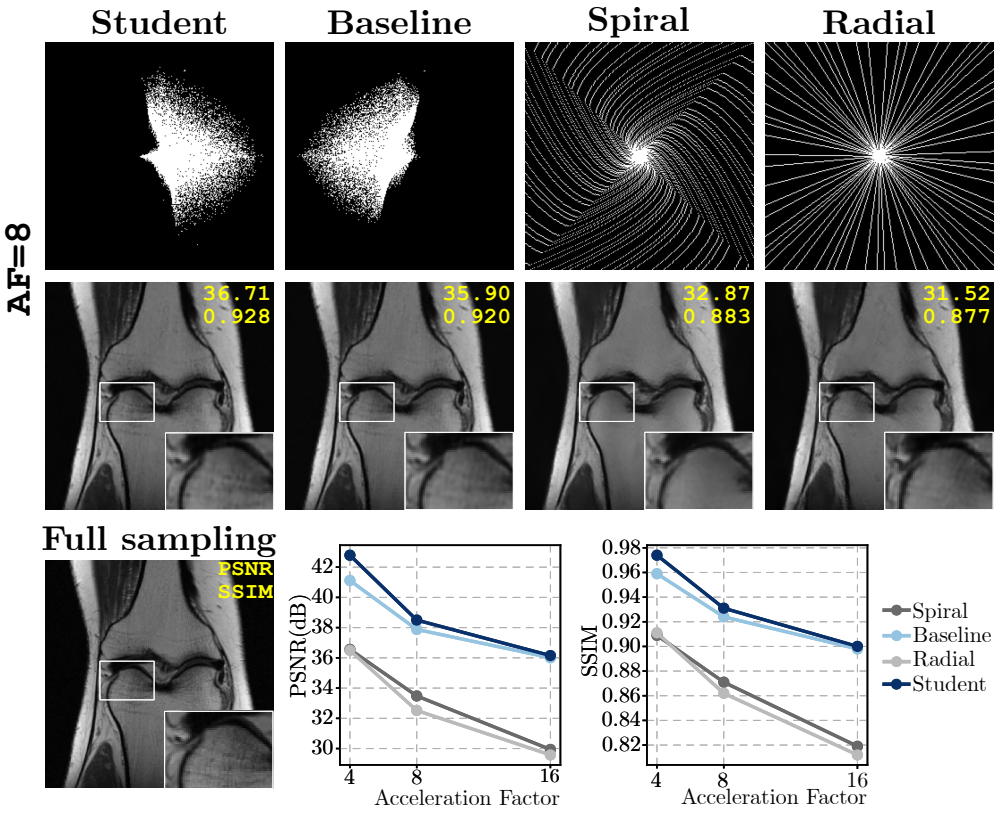


Table 1
Reconstruction performance of student and baseline MRI systems for $AF \in \{4, 8, 16\}$. The teacher system has an acceleration factor AF_t that is one unit lower than the corresponding student ($AF_t = AF_s - 1$). Best results are shown in bold

AF_s	Teacher ($AF_t = AF_s - 1$)		Student		Baseline	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
4	45.48	0.985	$\mathbf{42.82 \pm 0.12}$	$\mathbf{0.974 \pm 0.0004}$	41.52 ± 0.28	0.962 ± 0.0025
8	39.07	0.938	$\mathbf{38.62 \pm 0.06}$	$\mathbf{0.932 \pm 0.0006}$	38.26 ± 0.22	0.928 ± 0.0024
16	36.57	0.905	$\mathbf{36.26 \pm 0.17}$	$\mathbf{0.901 \pm 0.0022}$	35.96 ± 0.22	0.896 ± 0.0027

To evaluate the improvements in the design of the encoder, the optimized $k-$space undersampling masks for the baseline and student systems were extracted, fixed its weights, and used to reconstruct magnetic resonance images from the undersampled measurements using a reconstruction neural network with the same architecture as the employed U-Net. Additionally, the reconstruction performance of the recovery network using the student's learned $k-$space undersampling masks was compared to traditional undersampling mask patterns, such as spiral and radial, from Liu and Saloner (2014). Figure 7 presents visual results of the reconstruction performance for the trained neural network with the fixed learned masks of the student and baseline, as well as with the spiral and radial patterns, all with $AF = 8$. These results show that the network trained with the student's learned mask outperforms the baseline as well as the spiral and radial patterns. Zoomed-in results reveal that the network trained with the student's mask captures finer high-frequency features, leading to improved reconstruction. Additionally, a plot for $AF \in \{4, 8, 16\}$ demonstrates that the network trained with the student mask consistently outperforms the networks trained with the baseline, spiral, and radial masks across all acceleration factors.

Figure 7. Comparison of the student MRI system with the baseline and common k-space undersampling masks (spiral and radial). On the upper rows, a visual comparison for $AF = 8$ is presented, with PSNR (dB) and SSIM metrics displayed in the upper-right corner of each reconstruction. On the right bottom, a performance plot comparison of the student, baseline, radial and spiral masks with $AF \in \{4, 8, 16\}$ is shown.



6.2. Single-pixel camera

Training details: The FashionMNIST dataset Xiao et al. (2017) was employed. It consists of $60,000$ training and $10,000$ testing images of size 28×28 containing 10 classes of clothing. The training dataset was divided into $50,000$ images for training and $10,000$ for validation. All images were resized to 32×32. The training was performed for 50 epochs using the Adam optimizer Kingma (2014) with a learning rate of 5×10^{-4} and a batch size

of $B = 64$ images.

We compared the proposed KD approach with the traditional E2E optimization scheme for designing SPC systems. All teacher systems were previously trained with $\gamma_t = \gamma_s$ with real-valued CAs. Five realizations of the student and baseline models were evaluated, with average and standard deviation reported. An additional comparison was made using the traditional Hadamard basis for the CA patterns Zhang et al. (2017), with the computational decoder trained accordingly. Table 2 shows the results, where the student system outperforms both the baseline (by up to 1.05 (dB)) and the Hadamard basis across all compression ratio configurations. Additionally, the student models, in almost all cases, demonstrate greater stability than the baseline, showing lower standard deviations in the reconstruction metrics.

To evaluate the improvement in the codification of the CAs learned by the student, we present the mutual coherence $\mu(\mathbf{A}) = \max_{i \neq j} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle|$ of the student and baseline forward model matrices across all evaluated compression ratios in Figure 8a. These results show that the columns of $\mathbf{A}_{\mathbf{\Phi}_s^\star}$ (the student's sensing matrix) are less correlated with each other compared to the columns of $\mathbf{A}_{\mathbf{\Phi}^\star}$ (the baseline's sensing matrix). This indicates that the student can collect a more diverse set of measurements than the baseline. Additionally, Figure 8b shows the distribution of the normalized singular values of the student and baseline forward model operators. The student's matrix exhibits larger minimal singular values, leading to a lower condition number $\kappa(\mathbf{A}) = \frac{\sigma_{\max}(\mathbf{A})}{\sigma_{\min}(\mathbf{A})}$ compared to the baseline, showing that the student's system is more stable and robust than the baseline. Furthermore, Figure 9 provides visual results comparing sections of the Gram matrices $\mathbf{A}_{\mathbf{\Phi}_s^\star}^\top \mathbf{A}_{\mathbf{\Phi}_s^\star}$ for the student and $\mathbf{A}_{\mathbf{\Phi}^\star}^\top \mathbf{A}_{\mathbf{\Phi}^\star}$ for

Table 2
Reconstruction performance of student and baseline SPC systems for $\gamma \in \{0.05, 0.1, 0.2, 0.3, 0.4\}$. *The teacher system has the same compression ratios as the student* $\gamma_t = \gamma_s$, *however, it uses real-valued coded apertures. Best results are shown in bold*

γ	Teacher		Student		Baseline		Hadamard	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
0.05	25.59	0.915	**24.50 ± 0.04**	**0.894 ± 0.001**	24.03 ± 0.03	0.887 ± 0.002	18.80 ± 0.03	0.745 ± 0.004
0.1	27.84	0.944	**26.07 ± 0.04**	**0.920 ± 0.001**	25.53 ± 0.09	0.907 ± 0.009	19.78 ± 0.03	0.793 ± 0.006
0.2	30.90	0.972	**27.93 ± 0.06**	**0.945 ± 0.001**	27.06 ± 0.08	0.936 ± 0.002	22.60 ± 0.03	0.867 ± 0.013
0.3	33.29	0.978	**29.11 ± 0.08**	**0.955 ± 0.002**	28.06 ± 0.21	0.947 ± 0.003	24.95 ± 0.05	0.911 ± 0.006
0.4	34.57	0.987	**29.54 ± 0.06**	**0.961 ± 0.001**	28.81 ± 0.2	0.955 ± 0.002	26.83 ± 0.08	0.938 ± 0.004

the baseline. The student's Gram matrix demonstrates stronger linear independence due to its off-diagonal elements having lower values compared to the baseline's Gram matrix. This suggests that the student's matrix provides more distinct and non-redundant information, contributing to improved performance in the reconstruction due to measurement diversity and stability of the sensing matrix.

Figure 9. Visualization of a 256×256 section of the $\mathbf{A}_{\mathbf{\Phi}}^{\top}\mathbf{A}_{\mathbf{\Phi}}$ matrices for the student's SPC and the baseline's sensing matrices.

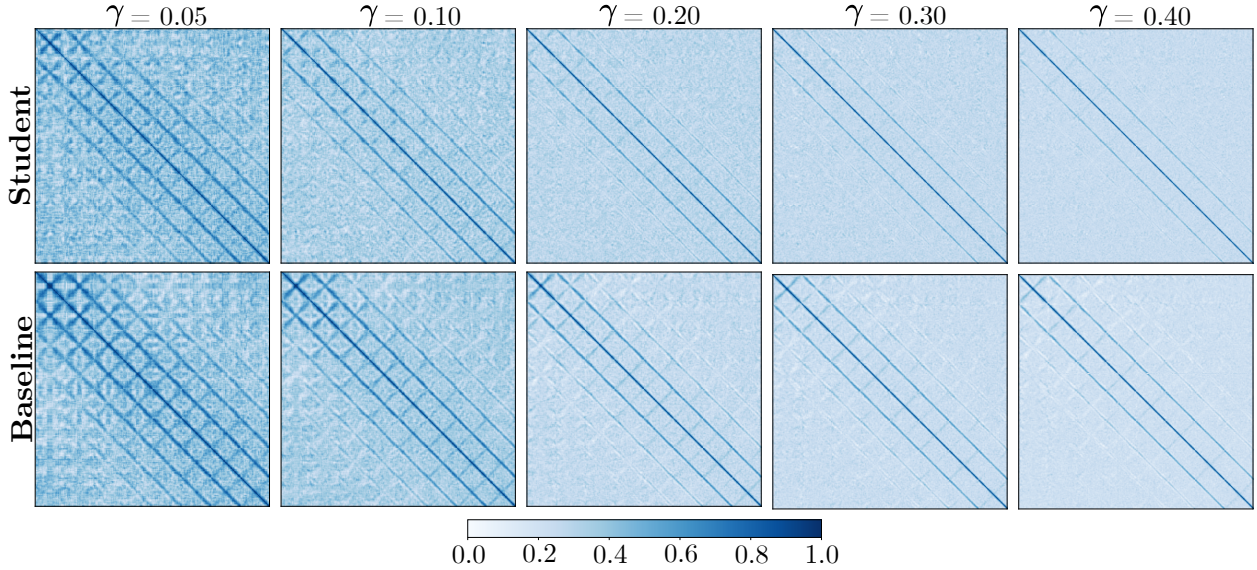
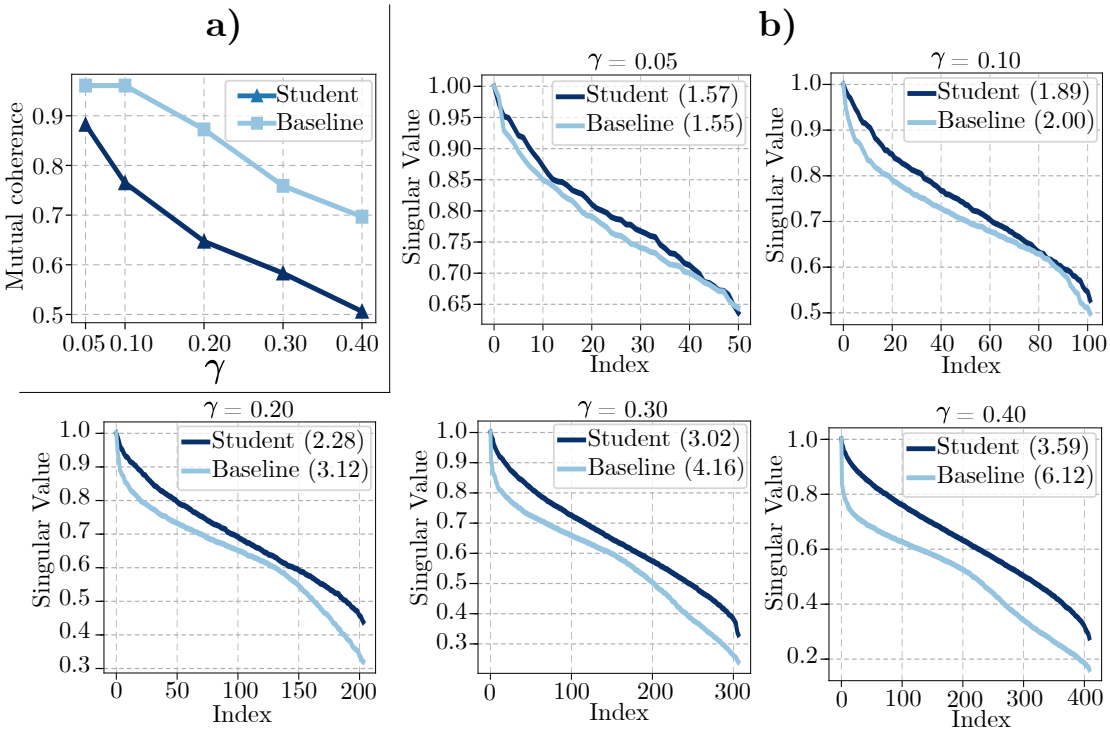


Figure 8. (a) Mutual coherence of the student and baseline SPC sensing matrices for various compression ratios. (b) Distribution of normalized singular values for the student and baseline SPC matrices at different compression ratios, with their condition numbers indicated next to the respective labels.



6.3. Single disperser coded aperture snapshot spectral imager

Training details: The ARAD-1K dataset Arad et al. (2022) was used for multi-spectral image reconstruction. It contains 950 images, each of size $31 \times 512 \times 512$ in the 400-700 nm range. The images were resized to 256×256 and eight equidistant spectral bands were selected from the original 31 bands. Four non-overlapping patches were extracted from each image, resulting in $3,600$ patches for training, 125 for validation, and 75 for testing, with each patch sized $8 \times 128 \times 128$. The AdamW Loshchilov (2017) optimizer with a weight decay of 1×10^{-2} was employed. The teacher, student, and baseline systems were trained for

500 epochs with a learning rate of 5×10^{-4}, using a batch size of $B = 32$.

The proposed approach was compared to traditional E2E optimization for a SD-CASSI system with a binary-valued CA and one-snapshot configuration. The teacher system was pretrained using a real-valued CA with the same one-snapshot configuration. Five realizations of the student and baseline models were conducted, with the average and standard deviation reported. Additionally, we compared the reconstruction performance of the network using a SD-CASSI with a Blue Noise CA Correa et al. (2016) with one snapshot. Table 3 summarizes the results, demonstrating that the student model outperforms the baseline (by up to 1.53 dB in PSNR) and Blue Noise CA (by up to 1.56 (dB) in PSNR).

Figure 10 presents visual reconstructions of all spectral bands, comparing the proposed KD approach with E2E optimization and the use of a Blue Noise CA as a fixed encoder. The results are shown in terms of PSNR and SSIM. The results demonstrate that the student model consistently outperforms both the baseline and Blue Noise CA across all bands, and in nearly all bands, it even surpasses the teacher model. For the multi-spectral image, the student achieves a PSNR of 40.90 (dB), surpassing the teacher's 40.30 (dB). The student also outperforms the baseline and Blue Noise CA, with improvements of 1.47 (dB) over the baseline and 2.04 (dB) over the Blue Noise CA. Also, in Figure 11 spectral profiles were evaluated at the RGB ground truth image point. The Spectral Angle Mapper (SAM) metric Yuhas et al. (1992) was used to quantify spectral similarity. The results show that the student system achieves a SAM score of 0.0344, outperforming the baseline score of 0.0433 and the Blue Noise CA score of 0.0617 (lower values indicate better similarity).

Table 3
Reconstruction performance of the student, baseline, Blue Noise CA, and teacher SD-CASSI systems. The baseline, student, and Blue Noise models use a binary-valued CA, while the teacher uses a real-valued CA.

	PSNR ↑	SSIM ↑
Teacher	37.44	0.965
Student	$\mathbf{36.89 \pm 0.01}$	$\mathbf{0.958 \pm 0.001}$
Baseline	35.36 ± 0.21	0.943 ± 0.002
Blue Noise	35.33 ± 0.03	0.942 ± 0.001

Figure 10. Reconstruction performance of the SD-CASSI system, reported in PSNR (dB) and SSIM, for the teacher, student, baseline, and Blue Noise models. Results are shown for all eight spectral bands, detailing the PSNR and SSIM for each band. The overall performance across all eight spectral bands is: teacher (PSNR: 40.30 (dB), SSIM: 0.971), student (PSNR: 40.90 (dB), SSIM: 0.971), baseline (PSNR: 39.43 (dB), SSIM: 0.965), and Blue Noise (PSNR: 38.86 (dB), SSIM: 0.960).

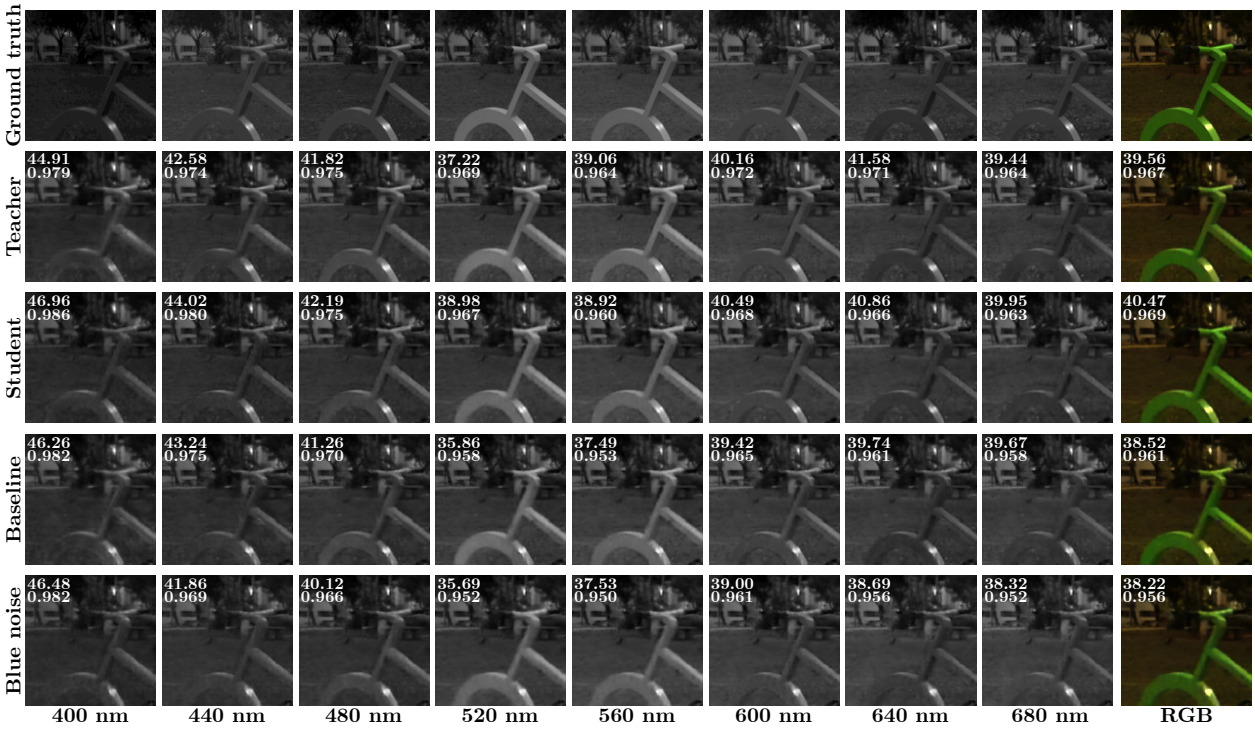
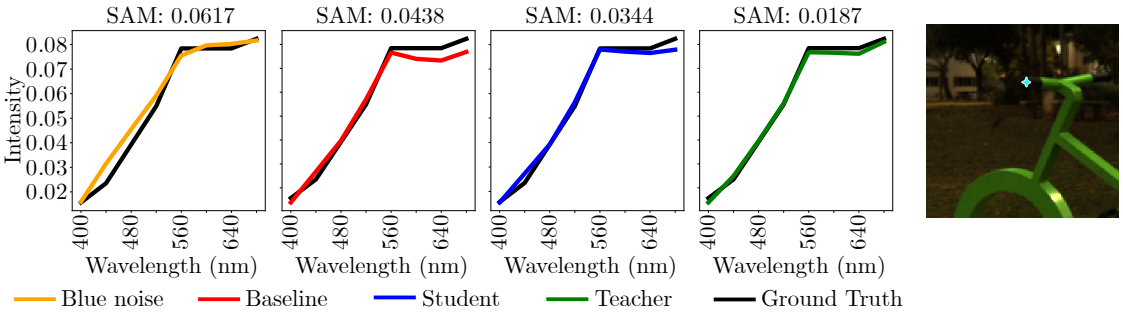


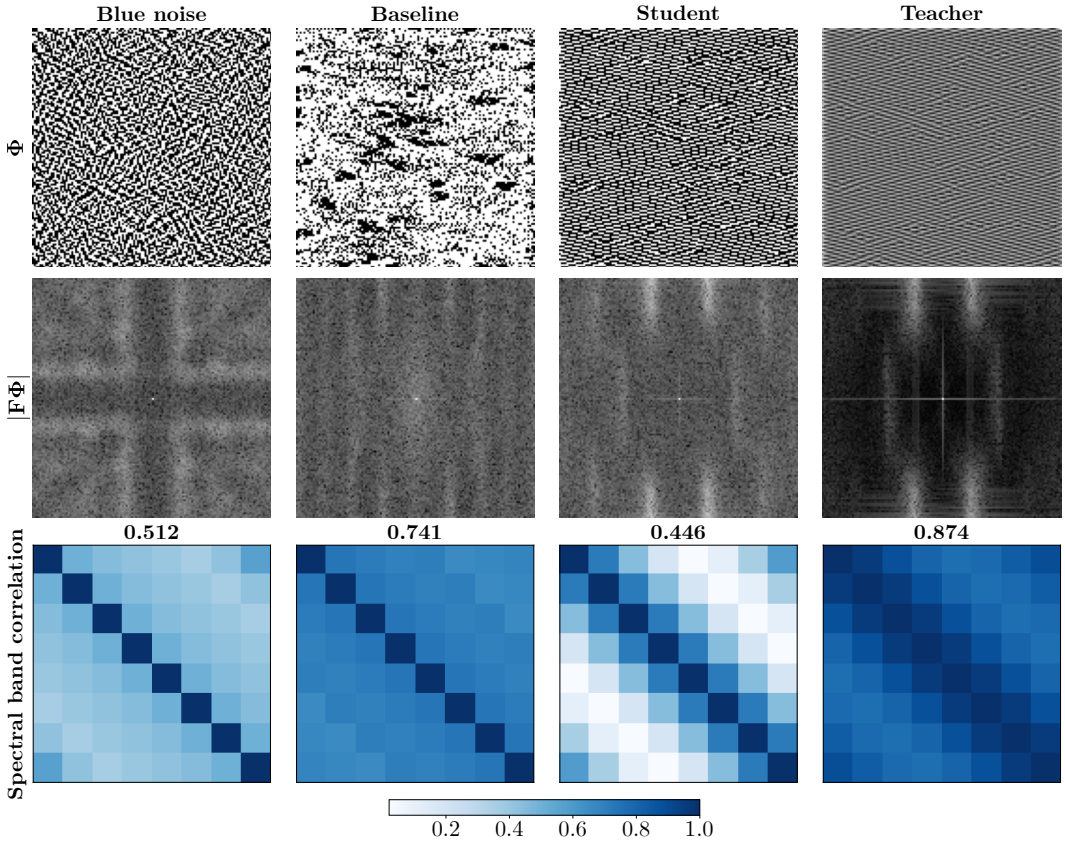
Figure 11. Spectral profiles for a point in the RGB ground truth image are presented, along with the SAM metric.



To evaluate the improvements in the design of the encoder for the SD-CASSI system using the proposed approach, we computed the spectral band correlation matrix $\mathbf{G} \in \mathbb{R}^{L\times L}$. This matrix quantifies the correlation between the spectral bands, where low correlation values indicate incoherent spectral sampling. The matrix is obtained by decomposing the SD-CASSI sensing matrix $\mathbf{A_\Phi} \in \mathbb{R}^{N(M+L-1)\times NML}$ into L blocks, such that $\mathbf{A_\Phi} = [\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_L]$, where each $\mathbf{A}_i \in \mathbb{R}^{N(M+L-1)\times NM}$ corresponds to the sensing matrix for the i-th spectral band, containing the diagonalized CA shifted by $i \cdot N$ positions. The (i, j)-th entry of the spectral band correlation matrix $\mathbf{G}$ is computed as the inner product of the submatrices $\mathbf{A}_i$ and $\mathbf{A}_j$, given by $\mathbf{G}_{i,j} = \mathrm{Tr}(\mathbf{A}_i^\top \mathbf{A}_j)$. Additionally, we computed the Fourier transform magnitude of the CAs $\mathbf{F\Phi}$. Figure 12 illustrates both the spectral band correlation matrix and the Fourier transform magnitude metrics for the baseline, teacher, student, and Blue Noise CAs. The results show that the teacher system focuses on very specific low and high frequencies, discarding the rest of the spectrum. The student system displays a similar frequency response to the teacher but with a broader emphasis across

the spectrum, particularly enhancing both low and high frequencies, while also prioritizing certain regions. The baseline CA shows a more uniform behavior, with higher-frequency details present but not as dominantly represented as in the teacher or student CAs. The Blue Noise CA provides a more uniform encoding of both low and high frequencies. However, it's not as high-frequency focused as a teacher, nor as low-frequency concentrated as student or baseline. Furthermore, the spectral band correlation matrix results show that the student CA exhibits lower coherence than the baseline and Blue Noise CAs, indicating that it is less spectrally correlated. This lower correlation suggests that the student CA acquires more diverse and less spectrally redundant measurements compared to the other two. Additionally, the student CA's average correlation of 0.446 is lower than that of the Baseline (0.741) and Blue Noise (0.512). The higher average spectral band correlation seen in the teacher system can be attributed to the relaxation of the binary constraint. In other words, allowing the teacher CA to use real-valued intensity levels, rather than strictly binary values, leads to measurements that are more closely related spectrally.

Figure 12. The first row displays the employed Blue Noise CA along with the learned CA of the baseline, student, and teacher. The second row presents the magnitude of the Fourier Transform for the Blue Noise CA and the learned CA of the baseline, student, and teacher. The third row shows the spectral band correlation matrix for the Blue Noise CA and the learned CA of the baseline, student, and teacher, with the average spectral band correlation displayed above each matrix.



7. Ablation Studies

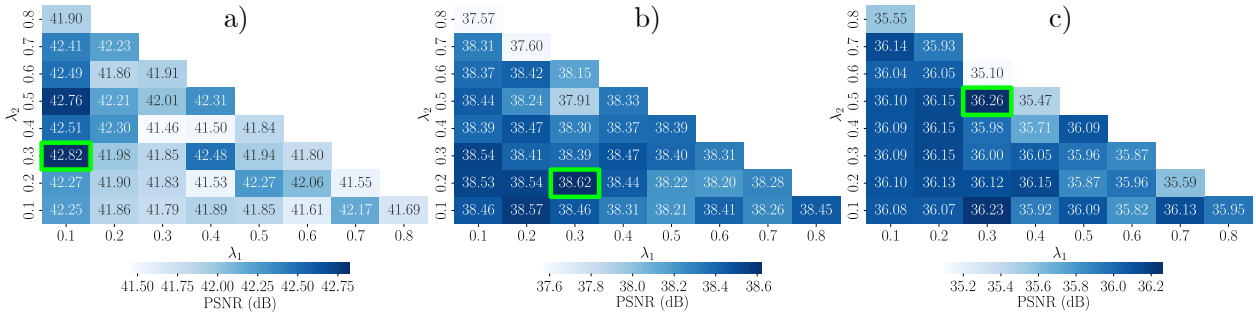
This chapter presents ablation studies conducted to determine the optimal hyperparameters for the knowledge transfer loss functions (λ_1, λ_2), evaluate the noise robustness of the proposed approach, identify the best teacher model, assess the contribution of each knowledge transfer loss function ($\mathcal{L}_{\text{ENC}}$ and $\mathcal{L}_{\text{DEC}}$), and analyze computational requirements during the training phase.

7.1. Hyperparameter tuning

This section presents the results of the hyperparameter tuning for the regularization terms of the KD loss function (λ_1, λ_2). A grid search was conducted for each CI system, evaluating $(\lambda_1, \lambda_2) \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\} \times \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$ with the constraint that $\lambda_1 + \lambda_2 \leq 1$ to prevent negative values for λ_3.

Figure 13 shows the results of this search for MRI, with the tuning performed for each acceleration factor of the student, i.e., $AF_s \in \{4, 8, 16\}$, where the teacher model used $AF_t = AF_s - 1$ for each corresponding AF_s. We observed the following trends: for $AF_s = 4$, the optimal set was $\{\lambda_1 = 0.1, \lambda_2 = 0.3\}$ with $\lambda_3 = 0.6$, indicating that the encoder loss function $\mathcal{L}_{\text{ENC}}$ plays a dominant role, with the student's performance heavily relying on mimicking the teacher's undersampling mask. For $AF_s = 8$, the optimal set was $\{\lambda_1 = 0.3, \lambda_2 = 0.2\}$ with $\lambda_3 = 0.5$, while for $AF_s = 16$, the optimal set was $\{\lambda_1 = 0.3, \lambda_2 = 0.5\}$ with $\lambda_3 = 0.2$. As the acceleration factor increases, the number of available points for subsampling the $k-$space decreases, shifting the focus of the loss function more towards knowledge transfer in the

Figure 13. Grid search for hyperparameter tuning of λ_1 and λ_2 for MRI students with acceleration factors a) $AF_s = 4$, b) $AF_s = 8$, and c) $AF_s = 16$, using a teacher model with $AF_t = AF_s - 1$ for each case. Results are reported in PSNR.



decoder and less on the encoder.

For SPC, the optimal values were found to be $\lambda_1 = \lambda_2 = 0.1$, leading to $\lambda_3 = 0.8$. This configuration indicated that the student's performance was primarily influenced by minimizing the encoder loss by aligning the student's gram matrix of the forward operator before binarization $\mathbf{A}_{\mathbf{W}_s}^\top \mathbf{A}_{\mathbf{W}_s}$ with the teacher's gram matrix $\mathbf{A}_{\mathbf{\Phi}_t^\star}^\top \mathbf{A}_{\mathbf{\Phi}_t^\star}$.

For the SD-CASSI system, a similar behavior was observed as with the SPC. The encoder loss function played a more significant role in the performance than the other losses. However, the decoder loss did not contribute to improving performance as much as the encoder loss. The optimal configuration was found to be $\lambda_1 = 0.1$, $\lambda_2 = 0$, and $\lambda_3 = 0.9$, with the system's improvement heavily relying on aligning the student's gram matrix of the CAs before binarization $\mathbf{W}_s^\top \mathbf{W}_s$ with the teacher's real-valued CAs gram matrix $\mathbf{\Phi}_t^{\star\top} \mathbf{\Phi}_t^\star$.

7.2. Noise robustness

Here, we evaluate the performance of the proposed approach in the presence of noise in the measurements. The analysis considers signal-to-noise ratios (SNRs) ranging from 20 dB to

40 dB, with increments of 5 dB. Table 4 presents the reconstruction results in terms of PSNR for MRI, showing that the student consistently outperforms the baseline across all noise configurations. For an acceleration factor of $AF = 4$, both the student and baseline models exhibit the lowest PSNR values. This is primarily because the computational decoder at this acceleration factor was trained with less severe degradations compared to the decoders at higher acceleration factors. Additionally, although the student shows improved performance under noisy conditions, the overall gains are limited. This can be attributed to the fact that the computational decoders were trained on noise-free data. The performance could potentially be enhanced by incorporating noise into the $k-$space measurements during the training phase.

Table 4
Reconstruction performance of the student and baseline models under varying noise levels for MRI systems in terms of PSNR. Additive Gaussian noise with SNR values ranging from 20 dB to 40 dB was added to the measurements.

		SNR (dB)				
	AF	20	25	30	35	40
Student	4	**35.53**	**39.08**	**41.26**	**42.27**	**42.64**
	8	**36.69**	**38.01**	**38.48**	**38.63**	**38.68**
	16	**35.81**	**36.21**	**36.34**	**36.39**	**36.40**
Baseline	4	35.28	38.80	40.73	41.53	41.81
	8	36.48	37.81	38.27	38.43	38.47
	16	35.57	35.94	36.05	36.09	36.10

Table 5 shows the reconstruction performance of the student and baseline models for SPC under noise conditions. In all cases, the student outperforms the baseline, except for

$\gamma = 0.05$ with a noise level of 20 dB. This exception may be attributed to the student having a slightly higher condition number (1.57) compared to the baseline (1.55), as shown in Figure 8b in the Simulations & Results Chapter.

Table 5
Reconstruction performance of the student and baseline models under varying noise levels for SPC systems in terms of PSNR. Additive Gaussian noise with SNR values ranging from 20 dB to 40 dB was added to the measurements.

		SNR (dB)				
	γ	20	25	30	35	40
Student	0.05	23.18	**24.07**	**24.40**	**24.51**	**24.54**
	0.1	**24.93**	**25.67**	**25.93**	**26.02**	**26.04**
	0.2	**26.92**	**27.66**	**27.91**	**27.98**	**28.01**
	0.3	**28.10**	**28.76**	**28.98**	**29.05**	**29.07**
	0.4	**28.59**	**29.22**	**29.43**	**29.50**	**29.52**
Baseline	0.05	**23.25**	23.80	23.99	24.05	24.07
	0.1	24.78	25.23	25.38	25.43	25.44
	0.2	26.33	26.77	26.90	26.95	26.96
	0.3	27.15	27.60	27.74	27.78	27.80
	0.4	28.26	28.69	28.83	28.87	28.89

Table 6 shows the reconstruction performance of the student and baseline models for the SD-CASSI system under noise conditions. In all noise configurations, the student significantly outperforms the baseline.

7.3. Selection of the best teacher

This section aims to determine which teacher CI system provides the most effective knowledge to the student. To achieve this, various teacher configurations were evaluated in the three CI systems.

Table 6
Reconstruction performance of the student and baseline models under varying noise levels for SD-CASSI systems in terms of PSNR. Additive Gaussian noise with SNR values ranging from 20 dB to 40 dB was added to the measurements.

SNR (dB)	Student	Baseline
20	**35.72**	34.68
25	**36.52**	35.30
30	**36.79**	35.50
35	**36.86**	35.55
40	**36.89**	35.57

In MRI, teachers with $AF_t \in \{3, 7, 15\}$ were used to distill students with $AF_s \in \{4, 8, 16\}$, under the condition that $AF_t < AF_s$. Table 7 presents the results, which show that the closer the teacher's acceleration factor is to the student's (while satisfying $AF_t < AF_s$), the better the student's performance. This is partly due to the encoder loss function, which encourages the student's $k-$space undersampling mask to imitate the teacher's mask. When the teacher and student have similar acceleration factors, the structure of the teacher's mask is easier for the student to replicate, as the available sampling points in $k-$space are more comparable. However, when the acceleration factors differ significantly, the difference in available sampling points increases, making it more challenging for the student to effectively imitate the teacher's mask.

In SPC, teachers with real-valued CAs with $\gamma_t \in \{0.05, 0.1, 0.2, 0.3, 0.4\}$ were used to distill students with binary-valued CAs with $\gamma_s \in \{0.05, 0.1, 0.2, 0.3, 0.4\}$ under the condition $\gamma_t \geq \gamma_s$. Table 8 shows the results, which indicate that the maximum reconstruction performance is obtained when $\gamma_s = \gamma_t$. This is due to the encoder loss function $\mathcal{L}_{\text{ENC}}$, which

Table 7
Performance of the student models when distilled from teacher systems with different acceleration factors ($AF_t \in \{3, 7, 15\}$) for $AF_s \in \{4, 8, 16\}$, with $AF_t < AF_s$. Results are reported in PSNR.

		AF_t			Baseline
		3	7	15	
AF_s	16	35.91	36.13	**36.41**	36.10
	8	38.31	**38.71**		38.11
	4	**42.82**			41.94
Teacher		45.48	39.07	36.57	

minimizes the discrepancy between the Gram matrices of the teacher's and student's forward models. When $\gamma_s = \gamma_t$, the structure of the teacher's and student's sensing matrices is more closely aligned, making it easier for the student to approximate the teacher's forward model. However, when $\gamma_t > \gamma_s$, the teacher has access to a greater number of effective measurements due to the higher γ_t. This leads to a Gram matrix for the teacher that represents a richer set of measurement correlations, which the student, constrained by fewer sampling points, struggles to replicate.

In the SD-CASSI system, two teacher configurations were evaluated. The first configuration uses a one-snapshot SD-CASSI with a real-valued CA, while the second employs a two-snapshot SD-CASSI with real-valued CAs. Table 9 presents the PSNR results obtained by distilling knowledge from these two teacher configurations into a student with a one-snapshot setup and a binary-valued CA. The results reveal a behavior similar to that observed in SPC: the student achieves maximum performance when its number of snapshots

Table 8
Ablation study using various real-valued CA SPC teachers to optimize different binary-valued students under the criterion $\gamma_s \leq \gamma_t$. Results are reported in PSNR.

		γ_t					Baseline
		0.05	0.1	0.2	0.3	0.4	
	0.4					**29.46**	28.62
	0.3				**29.07**	28.44	28.13
γ_s	0.2			**27.85**	27.18	26.73	27.07
	0.1		**25.99**	24.02	23.67	23.98	25.70
	0.05	**24.46**	22.76	21.83	21.55	22.04	24.04
Teachers		25.59	27.84	30.90	33.29	34.57	

matches that of the teacher (one snapshot). This can be attributed to the encoder loss function, $\mathcal{L}_{\text{ENC}}$, which minimizes the discrepancy between the student's and teacher's CA Gram matrices. When the number of snapshots is the same for both the student and the teacher, the structure of the teacher's Gram matrix closely resembles what the student can replicate, facilitating the learning process. However, when the teacher uses more snapshots than the student, the student struggles to replicate the teacher's Gram matrix, resulting in reduced performance.

7.4. Knowledge distillation loss function ablation studies

In this section, the contributions of the different components of the KD loss function in (5) are evaluated across the three CI systems. Table 10 presents the results obtained using various configurations of the KD loss function. The findings indicate that each loss term contributes individually to improved reconstruction performance compared to the baseline. Notably, the encoder loss function provides the most significant improvement, as it directly

Table 9
Performance of student models (SD-CASSI with one snapshot and binary-valued CA) distilled from teacher systems configured with one or two snapshots and real-valued CAs. Results are reported in PSNR.

	Teacher Snapshots		Baseline
	1	2	
Student 1 snapshot	**36.90**	36.35	35.42
Teacher	37.44	38.71	

aligns the student's forward model with the teacher's. Additionally, in almost all cases, combining the two distillation loss terms with the task loss function yields the greatest improvements, highlighting the benefits of jointly distilling both the encoder and decoder components.

7.5. Computational requirements comparison

In this section, we compare the computational costs of the proposed KD method and the baseline for optimizing the three CI systems. All experiments used a NVIDIA RTX 3080 TI GPU. Table 11 presents the results comparing the proposed KD method with the baseline E2E method, focusing on GPU memory usage and training time for the CI systems: MRI with $AF = 4$, SPC with $\gamma = 0.4$, and SD-CASSI with one snapshot. These results reveal that while the KD method requires more resources than the baseline E2E method during training, the increases are relatively modest across all three CI systems. For MRI, the KD approach adds 13 minutes to the training time and requires 1,586 MB more memory. In SPC, the increase is 8 minutes in training time and 82 MB in memory. The largest difference is

Table 10
Ablation study on the KD loss function components ($\mathcal{L}_{ENC}$, $\mathcal{L}_{DEC}$) across three CI systems. Results are presented in terms of PSNR and SSIM.

MRI				
$\mathcal{L}_{\text{ENC}}$	✓	✓	✘	✘
$\mathcal{L}_{\text{DEC}}$	✘	✓	✓	✘
PSNR ↑	41.64	**42.56**	41.38	41.25
SSIM ↑	0.964	**0.973**	0.961	0.960
SPC				
$\mathcal{L}_{\text{ENC}}$	✓	✓	✘	✘
$\mathcal{L}_{\text{DEC}}$	✘	✓	✓	✘
PSNR ↑	29.40	**29.52**	28.78	28.62
SSIM ↑	0.956	**0.961**	0.956	0.955
SD-CASSI				
$\mathcal{L}_{\text{ENC}}$	✓	✓	✘	✘
$\mathcal{L}_{\text{DEC}}$	✘	✓	✓	✘
PSNR ↑	**36.90**	36.66	35.86	35.42
SSIM ↑	**0.958**	0.957	0.949	0.945

observed for SD-CASSI, where KD adds 50 minutes and 5,898 MB in memory usage, due the the features map of spectral images being significantly larger than the feature maps of one or two channel images. Despite these increases, the KD proposed approach is justified by the significant improvements in reconstruction performance and encoder design, as demonstrated in previous sections.

Additionally, the inference time remains the same for both the baseline and the proposed method because, after the knowledge transfer process during training, the teacher model is no longer needed. The student model, which has already learned the necessary knowledge from the teacher, performs inference independently. As a result, there are no additional

Table 11
Comparison of computational cost between the proposed method (KD) and the baseline (E2E) for the three CI systems: MRI with AF = 8, SPC with $\gamma = 0.4$, and SD-CASSI with one snapshot. The metrics include execution time and memory usage.

System	Training Time		Memory Usage (MB)	
	E2E	KD	E2E	KD
MRI	22m 18s	35m 13s	4,084	5,670
SPC	12m 12s	20m 22s	1,280	1,362
SD-CASSI	2h 13m 20s	3h 3m 4s	2,405	8,303

computational costs during inference.

8. Conclusions

This research introduces a novel, general-purpose approach for designing CI systems based on KD, demonstrating competitive performance against traditional E2E optimization. We validated this method through extensive experiments on three representative CI systems—MRI, the SPC, and the SD-CASSI. The results show that by relaxing and solving the original optimization problem, the teacher model effectively transfers valuable knowledge via encoder and decoder loss functions, significantly enhancing the student model's recovery performance. This approach achieves gains of up to 1.47 dB in MRI, 1.05 dB in SPC, and 1.53 dB in SD-CASSI.

Additionally, improvements in encoder design were observed, including better condition numbers, mutual coherence, and linear independence in SPC, better spectral band correlation in SD-CASSI, and enhanced reconstruction performance in MRI when incorporating the learned undersampling mask as a fixed encoder within reconstruction networks.

Moreover, validation experiments confirm the robustness of the proposed method under noisy measurement conditions. Ablation studies further suggest that a teacher system with an encoder structure similar to the student's—having a comparable number of measurements or a similar codification basis—provides more effective guidance.

9. Discussion & future work

The proposed approach is designed as an alternative to traditional E2E optimization for the design of any CI system by reinterpreting the concept of KD within CI systems. Instead of using a large and complex neural network as a teacher, it uses a less constrained, high-performance CI system, and the student corresponds to a more constrained CI system. In this work, we validated it on three representative CI systems: MRI, the SPC, and the SD-CASSI. Future work can address limitations of the proposed approach, such as the need for hyperparameter tuning of the KD loss functions, as these are not the same across all CI modalities, and determining the criteria for selecting an optimal teacher to guide the student. Further research could also explore the design of other CI systems, such as seismic imaging, phase retrieval, and diffractive optical imaging systems. Additionally, while only a few relaxations of the teacher's encoder were explored, future research could investigate less conventional parameterization found useful in neural networks and inverse problems such as complex Lee et al. (2022) or hypercomplex numbers Bojesomo et al. (2024); Jacome et al. (2024), or employ a different CI system as the teacher, distinct from the student. Furthermore, this work addresses the image reconstruction task; future work could explore other tasks, such as segmentation, classification, and depth estimation. Additionally, this approach can be useful for designing recovery-only schemes without designing the CE, which can be useful for a wide range of imaging inverse problems such as super-resolution, computed tomography, (non-) blind deconvolution, and denoising Ongie et al. (2020).

References

Alkhulaifi, A., Alsahli, F., and Ahmad, I. (2020). Knowledge distillation in deep learning and its applications. *PeerJ Computer Science*, 7:1–24.

Arad, B., Timofte, R., Yahel, R., Morag, N., Bernat, A., Cai, Y., Lin, J., Lin, Z., Wang, H., Zhang, Y., Pfister, H., Van Gool, L., Liu, S., Li, Y., Feng, C., Lei, L., Li, J., Du, S., Wu, C., Leng, Y., Song, R., Zhang, M., Song, C., Zhao, S., Lang, Z., Wei, W., Zhang, L., Dian, R., Shan, T., Guo, A., Feng, C., Liu, J., Agarla, M., Bianco, S., Buzzelli, M., Celona, L., Schettini, R., He, J., Xiao, Y., Xiao, J., Yuan, Q., Li, J., Zhang, L., Kwon, T., Ryu, D., Bae, H., Yang, H.-H., Chang, H.-E., Huang, Z.-K., Chen, W.-T., Kuo, S.-Y., Chen, J., Li, H., Liu, S., Sabarinathan, S., Uma, K., Bama, B. S., and Roomi, S. M. M. (2022). Ntire 2022 spectral recovery challenge and data set. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 862–880.

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

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

Arguello, H., Bacca, J., Kariyawasam, H., Vargas, E., Marquez, M., Hettiarachchi, R., Gar-

cia, H., Herath, K., Haputhanthri, U., Ahluwalia, B. S., So, P., Wadduwage, D. N., and Edussooriya, C. U. (2023). Deep optical coding design in computational imaging: A data-driven framework. *IEEE Signal Processing Magazine*, 40(2):75–88.

Bacca, J., Correa, C. V., Vargas, E., Castillo, S., and Arguello, H. (2019). Compressive classification from single pixel measurements via deep learning. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6.

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

Bacca, J., Gelvez-Barrera, T., and Arguello, H. (2021). Deep coded aperture design: An end-to-end approach for computational imaging tasks. *IEEE Transactions on Computational Imaging*, 7:1148–1160.

Bahadir, C. D., Wang, A. Q., Dalca, A. V., and Sabuncu, M. R. (2020). Deep-learning-based optimization of the under-sampling pattern in mri. *IEEE Transactions on Computational Imaging*, 6:1139–1152.

Bernet, S. (2018). Zoomable telescope by rotation of toroidal lenses. *Applied Optics*, 57(27):8087.

Bhandari, A., Kadambi, A., and Raskar, R. (2022). *Computational Imaging*. MIT Press.

Bojesomo, A., Liatsis, P., and Al Marzouqi, H. (2024). Deep hypercomplex networks for

spatiotemporal data processing: Parameter efficiency and superior performance [hypercomplex signal and image processing]. *IEEE Signal Processing Magazine*, 41(3):101–112.

Chen, W., Peng, L., Huang, Y., Jing, M., and Zeng, X. (2021). Knowledge distillation for u-net based image denoising. In *2021 IEEE 14th International Conference on ASIC (ASICON)*, pages 1–4.

Cho, J. H. and Hariharan, B. (2019). On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Chu, J., Du, C., Lin, X., Zhang, X., Wang, L., Zhang, Y., and Wei, H. (2025). Highly accelerated mri via implicit neural representation guided posterior sampling of diffusion models. *Medical Image Analysis*, 100:103398.

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

Courbariaux, M., Bengio, Y., and David, J.-P. (2015). Binaryconnect: Training deep neural networks with binary weights during propagations. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Duarte, M. F., Davenport, M. A., Takhar, D., Laska, J. N., Sun, T., Kelly, K. F., and Baraniuk, R. G. (2008). Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, 25(2):83–91.

Fang, H., Long, Y., Hu, X., Ou, Y., Huang, Y., and Hu, H. (2023). Dual cross knowledge distillation for image super-resolution. *Journal of Visual Communication and Image Representation*, 95:103858.

Gao, Q., Zhao, Y., Li, G., and Tong, T. (2019). Image super-resolution using knowledge distillation. In Jawahar, C. V., Li, H., Mori, G., and Schindler, K., editors, *Computer Vision – ACCV 2018*, pages 527–541, Cham. Springer International Publishing.

Geethanath, S., Reddy, R., Konar, A. S., Imam, S., Sundaresan, R., D. R., R. B., and Venkatesan, R. (2013). Compressed sensing mri: A review. *Critical Reviews in Biomedical Engineering*, 41(3):183–204.

GharehMohammadi, F. and Sebro, R. A. (2024). Efficient health care: Decreasing mri scan time. *Radiology: Artificial Intelligence*, 6(3).

Gou, J., Yu, B., Maybank, S. J., and Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

He, Z., Dai, T., Lu, J., Jiang, Y., and Xia, S.-T. (2020). Fakd: Feature-affinity based knowledge distillation for efficient image super-resolution. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 518–522.

Henry, A. and Gonzalo, R. A. (2011). Code aperture optimization for spectrally agile compressive imaging. *J. Opt. Soc. Am. A*, 28(11):2400–2413.

Hernandez-Rojas, A. and Arguello, H. (2024). Design of undersampled seismic acquisition geometries via end-to-end optimization. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jacome, R., Gomez, P., and Arguello, H. (2023). Middle output regularized end-to-end optimization for computational imaging. *Optica*, 10(11):1421–1431.

Jacome, R., Mishra, K. V., Sadler, B. M., and Arguello, H. (2024). An invitation to hypercomplex phase retrieval: Theory and applications [hypercomplex signal and image processing]. *IEEE Signal Processing Magazine*, 41(3):22–32.

Karl, W. C., Fowler, J. E., Bouman, C. A., Çetin, M., Wohlberg, B., and Ye, J. C. (2023). The foundations of computational imaging: A signal processing perspective. *IEEE Signal Processing Magazine*, 40(5):40–53.

Kellman, M. R., Bostan, E., Repina, N. A., and Waller, L. (2019). Physics-based learned design: Optimized coded-illumination for quantitative phase imaging. *IEEE Transactions on Computational Imaging*, 5(3):344–353.

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

Knoll, F., Zbontar, J., Sriram, A., Muckley, M. J., Bruno, M., Defazio, A., Parente, M., Geras, K. J., Katsnelson, J., Chandarana, H., Zhang, Z., Drozdzalv, M., Romero, A., Rabbat, M., Vincent, P., Pinkerton, J., Wang, D., Yakubova, N., Owens, E., Zitnick, C. L., Recht, M. P., Sodickson, D. K., and Lui, Y. W. (2020). FastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning. *Radiol. Artif. Intell.*, 2(1):e190007.

Lee, C., Hasegawa, H., and Gao, S. (2022). Complex-valued neural networks: A comprehensive survey. *IEEE/CAA Journal of Automatica Sinica*, 9(8):1406–1426.

Li, L., Wang, L., Song, W., Zhang, L., Xiong, Z., and Huang, H. (2022). Quantization-aware deep optics for diffractive snapshot hyperspectral imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19780–19789.

Liang, J., editor (2024). *Coded Optical Imaging*. Springer International Publishing.

Liu, J., Anirudh, R., Thiagarajan, J. J., He, S., Mohan, K. A., Kamilov, U. S., and Kim, H. (2023). Dolce: A model-based probabilistic diffusion framework for limited-angle ct reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10498–10508.

Liu, J. and Saloner, D. (2014). Accelerated MRI with CIRcular cartesian UnderSampling (CIRCUS): a variable density cartesian sampling strategy for compressed sensing and parallel imaging. *Quant. Imaging Med. Surg.*, 4(1):57–67.

Lopez, J., Vargas, E., and Arguello, H. (2024). Depth estimation from a single optical encoded image using a learned colored-coded aperture. *IEEE Transactions on Computational Imaging*, 10:752–761.

Loshchilov, I. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lustig, M., Donoho, D., and Pauly, J. M. (2007). Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195.

Mait, J. N., Euliss, G. W., and Athale, R. A. (2018). Computational imaging. *Adv. Opt. Photon.*, 10(2):409–483.

Mao, T., Cuadros, A. P., Ma, X., He, W., Chen, Q., and Arce, G. R. (2018). Fast optimization of coded apertures in x-ray computed tomography. *Optics Express*, 26(19):24461–24478.

Metzler, C. A., Ikoma, H., Peng, Y., and Wetzstein, G. (2020). Deep optics for single-shot high-dynamic-range imaging. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1372–1382.

Miyamoto, K. (1961). The phase fresnel lens. *J. Opt. Soc. Am.*, 51(1):17–20.

Mosher, C., Li, C., Morley, L., Ji, Y., Janiszewski, F., Olson, R., and Brewer, J. (2014). Increasing the efficiency of seismic data acquisition via compressive sensing. *The Leading Edge*, 33(4):386–391.

Mosher, C. C., Li, C., Janiszewski, F. D., Williams, L. S., Carey, T. C., and Ji, Y. (2017). Operational deployment of compressive sensing systems for seismic data acquisition. *The Leading Edge*, 36(8):661–669.

Munoz, C., Fotaki, A., Botnar, R. M., and Prieto, C. (2022). Latest advances in image acceleration: All dimensions are fair game. *J Magn Reson Imaging*, 57(2):387–402.

Oemrawsingh, S. S. R., van Houwelingen, J. A. W., Eliel, E. R., Woerdman, J. P., Verstegen, E. J. K., Kloosterboer, J. G., and 't Hooft, G. W. (2004). Production and characterization of spiral phase plates for optical wavelengths. *Applied Optics*, 43(3):688.

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

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library.

Quan, Y., Chen, Z., Pang, T., and Ji, H. (2023). Unsupervised deep learning for phase retrieval via teacher-student distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):2128–2136.

Ramalli, A., Boni, E., Roux, E., Liebgott, H., and Tortoli, P. (2022). Design, implemen-

tation, and medical applications of 2-d ultrasound sparse arrays. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 69(10):2739–2755.

Razumov, A., Rogov, O., and Dylov, D. V. (2023). Optimal mri undersampling patterns for ultimate benefit of medical vision tasks. *Magnetic Resonance Imaging*, 103:37–47.

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. (2015). Fitnets: Hints for thin deep nets.

Ronneberger, O., Fischer, P., and Brox, T. (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation*, page 234–241. Springer International Publishing.

Sun, J., Su, Y., Zhang, H., Cheng, Z., Zeng, Z., Wang, Z., Chen, B., and Yuan, X. (2024). Snapcap: Efficient snapshot compressive video captioning.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tachella, J., Chen, D., Hurault, S., Terris, M., and Wang, A. (2023). DeepInverse: A deep learning framework for inverse problems in imaging.

Urrea, S., Jacome, R., Asif, M. S., Arguello, H., and Garcia, H. (2024). Dodo: Double doe optical system for multishot spectral imaging. *IEEE Journal of Selected Topics in Signal Processing*.

Vasanawala, S., Murphy, M., Alley, M., Lai, P., Keutzer, K., Pauly, J., and Lustig, M. (2011). Practical parallel imaging compressed sensing mri: Summary of two years of experience in accelerating body mri of pediatric patients. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1039–1043.

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

Wang, L. and Yoon, K.-J. (2022). Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3048–3068.

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

Weber, T., Ingrisch, M., Bischl, B., and Rügamer, D. (2024). Constrained probabilistic mask learning for task-specific undersampled mri reconstruction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7665–7674.

Weiss, T., Vedula, S., Senouf, O., Michailovich, O., Zibulevsky, M., and Bronstein, A. (2020). Joint learning of cartesian under sampling andre construction for accelerated mri. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8653–8657.

Wetzstein, G., Ozcan, A., Gigan, S., Fan, S., Englund, D., Soljačić, M., Denz, C., Miller, D. A. B., and Psaltis, D. (2020). Inference in artificial intelligence with deep optics and photonics. *Nature*, 588(7836):39–47.

Wu, Z. and Li, X. (2024). Adaptive knowledge distillation for high-quality unsupervised mri reconstruction with model-driven priors. *IEEE Journal of Biomedical and Health Informatics*, 28(6):3571–3582.

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms.

Xie, J., Gong, L., Shao, S., Lin, S., and Luo, L. (2023). Hybrid knowledge distillation from intermediate layers for efficient single image super-resolution. *Neurocomputing*, 554:126592.

Yiasemis, G., Sánchez, C. I., Sonke, J.-J., and Teuwen, J. (2024). On retrospective k-space subsampling schemes for deep mri reconstruction. *Magnetic Resonance Imaging*, 107:33–46.

Yim, J., Joo, D., Bae, J., and Kim, J. (2017). A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yuhas, R. H., Goetz, A. F., and Boardman, J. W. (1992). Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In *JPL,*

Summaries of the Third Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop.

Zaitsev, M., Maclaren, J., and Herbst, M. (2015). Motion artifacts in mri: A complex problem with many partial solutions. *Journal of Magnetic Resonance Imaging*, 42(4):887–901.

Zhang, Z., Wang, X., Zheng, G., and Zhong, J. (2017). Hadamard single-pixel imaging versus fourier single-pixel imaging. *Optics Express*, 25(16):19619–19639.

Zhu, H., Chen, Z., and Liu, S. (2023). Learning knowledge representation with meta knowledge distillation for single image super-resolution. *Journal of Visual Communication and Image Representation*, 95:103874.