A DEEP DISTILLATION ALGORITHM FOR NON-LINEAR GRADIENT PRECONDITIONING IN INVERSE PROBLEMS

YESID ROMARIO GUALDRÓN HURTADO
Systems Engineer

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTY OF PHYSICOMECHANICAL ENGINEERING
SCHOOL OF SYSTEMS ENGINEERING AND INFORMATICS
BUCARAMANGA
2025

A DEEP DISTILLATION ALGORITHM FOR NON-LINEAR GRADIENT PRECONDITIONING IN INVERSE PROBLEMS

YESID ROMARIO GUALDRÓN HURTADO

In fulfillment of the requirements for the degree of

Master in Systems Engineering and Informatics

Advisor:

Henry Arguello Fuentes

Ph.D. Electrical and Computer Engineering

UNIVERSIDAD INDUSTRIAL DE SANTANDER

FACULTY OF PHYSICOMECHANICAL ENGINEERING

SCHOOL OF SYSTEMS ENGINEERING AND INFORMATICS

MASTER DEGREE IN SYSTEMS ENGINEERING AND INFORMATICS

BUCARAMANGA

2025

CONTENT

LIST OF FIGURES

LIST OF TABLES

RESUMEN

TÍTULO: A DEEP DISTILLATION ALGORITHM FOR NON-LINEAR GRADIENT PRECONDITIONING IN INVERSE PROBLEMS [*]

AUTOR: YESID ROMARIO GUALDRÓN HURTADO [**]

PALABRAS CLAVE: Algoritmos de reconstrucción, preacondicionamiento del gradiente, destilación del conocimiento, problemas inversos.

DESCRIPCIÓN: Los algoritmos de reconstrucción que combinan optimización y regularización permiten la integración de los modelos físicos bien definidos con eliminadores de ruido para resolver problemas inversos en imagenología. Sin embargo, la solución del término de fidelidad de datos plantea desafíos significativos debido a la matriz de adquisición mal condicionada ocasionada por las restricciones físicas en el sistema de adquisición. Los algoritmos han adoptado técnicas de preacondicionamiento para abordar el mal condicionamiento, mejorando así la optimización del término de fidelidad y la velocidad de convergencia. No obstante, los diseños actuales del operador de preacondicionamiento (PO) se basan en la estructura de la matriz de adquisición o en diseños de extremo a extremo, lo que puede limitar el rendimiento, debido a que la estructura puede ser suboptima y por el desvanecimiento del gradiente, respectivamente. Por lo tanto, introducimos la destilación de conocimiento (KD) en algoritmos para diseñar un preacondicionamiento del gradiente no lineal (D^2GP) mediante la guía controlada de un algoritmo mejor condicionado. Se construyó un algoritmo maestro (TA) que emplea una matriz de adquisición simulada (virtual) con pocas restricciones físicas —solo factible en simulaciones—, lo que permite un alto rendimiento en la recuperación. El algoritmo estudiante (SA) utiliza una matriz de adquisición físicamente factible, que limita el rendimiento en la recuperación. El PO se diseña de tal manera que, al integrarse en el SA, puede alcanzar un rendimiento similar al del TA. Se diseñaron diferentes funciones de pérdida de destilación para transferir distintas propiedades del TA al SA. Se validó el diseño propuesto del PO en varias modalidades de imagenología, tales como la resonancia magnética, la cámara de un solo píxel y superresolución.

---

[*] Tesis de Maestría

[**] Faculty of Physicomechanical Engineering. School of Systems Engineering and Informatics. Director: Henry Arguello Fuentes.

ABSTRACT

TITLE: A Deep Distillation Algorithm for Non-linear Gradient Preconditioning in Inverse Problems [*]

AUTHOR: YESID ROMARIO GUALDRÓN HURTADO [**]

KEYWORDS: Recovery algorithms, gradient preconditioning, knowledge distillation, inverse problems.

DESCRIPTION: Recovery algorithms that combine optimization and regularization enable the integration of well-defined physical forward models with state-of-the-art denoisers to solve imaging inverse problems. However, solving the data fidelity term poses significant challenges due to the ill-conditioned sensing matrix caused by physical constraints in the acquisition system. Algorithms have adopted preconditioning techniques to address the ill-conditioning, enhancing the data fidelity optimization and the convergence speed. However, current designs for the preconditioning operator (PO) are often based on the structure of the sensing matrix or designed in an end-to-end manner, which may limit performance, due to suboptimal structure or the gradient vanishing, respectively. Thereby, knowledge distillation (KD) is introduced in algorithms to design a nonlinear gradient preconditioning (D^2GP) through the controlled guidance of a best-conditioned algorithm. A teacher algorithm (TA) was constructed, it employs a simulated (virtual) sensing matrix with few physical constraints—only feasible in simulations—allowing for high recovery performance. The student algorithm (SA) uses a physically feasible sensing matrix, which typically limits the recovery performance. The PO is designed so that when integrated into the SA, it can have a performance similar to the TA. Different distillation loss functions to transfer different properties of the TA to the SA were designed. The proposed PO design was validated in several imaging modalities such as magnetic resonance imaging, single-pixel camera, and super-resolution imaging.

---

[*] Master Thesis

[**] Faculty of Physicomechanical Engineering. School of Systems Engineering and Informatics. Advisor: Henry Arguello Fuentes.

Research Products

Contributions of the thesis

- A nonlinear gradient preconditioning through knowledge distillation is proposed, surpassing state-of-the-art preconditioning methods with fewer parameters by eliminating the dependence on signal dimensionality.
- Different distillation loss functions are proposed to match the student's and teacher's outputs and data fidelity gradients.
- A regularization function over the PO design is proposed based on the theoretical convergence rate of the gradient preconditioned student algorithm (GPSA) to accelerate its convergence.
- Improved recovery quality with fewer iterations, regardless of the condition number and the fidelity loss.
- The measurement augmentation methodology is proposed, and allows, through neural networks, to generate new measurements that improve the reconstruction quality.
- The integration of measurement augmentation methodology into algorithms, results in non-linear gradient preconditioning.
- The proposed methodologies are extensively validated in three imaging inverse problems and can be extended in other computational imaging and computer vision scenarios.

Publications

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

Journal papers:

1. **Romario Gualdrón-Hurtado**, Henry Arguello, and Jorge Bacca. **Deep Learned Non-Linear Propagation Model Regularizer for Compressive Spectral Imaging.** Published in *IEEE Transactions on Computational Imaging.* [1]

2. **Romario Gualdrón-Hurtado**, Roman Jacome, and Henry Arguello. **Measurement Completion for Inverse Problems.** Soon to be submitted to *IEEE Journal of Selected Topics in Signal Processing.*

3. **Romario Gualdrón-Hurtado**, Felipe B. Da Silva, Jorge Bacca, and Henry Arguello. **Patch-based Deep Coded Aperture Design for the Near-Infrared Spectral Range.** Soon to be submitted to *Applied Optics, Optica Publishing Group.*

Conference papers:

1. **Romario Gualdrón-Hurtado**, Roman Jacome, Leon Suarez, Emmanuel Martinez, and Henry Arguello. **Improving Compressive Imaging Recovery via Measurement Augmentation.** Presented in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2025).* [2]

---

[1] Romario Gualdrón-Hurtado, Henry Arguello, and Jorge Bacca. "Deep Learned Non-Linear Propagation Model Regularizer for Compressive Spectral Imaging". In: *IEEE Transactions on Computational Imaging* 10 (2024), pp. 1016–1025. DOI: 10.1109/TCI.2024.3422900.

[2] Romario Gualdrón-Hurtado et al. "Improving Compressive Imaging Recovery via Measurement Augmentation". In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2025, pp. 1–5. DOI: 10.1109/ICASSP49660.2025.10888734.

2. Emmanuel Martinez, Leon Suarez, **Romario Gualdrón-Hurtado**, Roman Jacome, and Henry Arguello. **Compressive Imaging Reconstruction via Conditional Diffusion Model With Augmented Measurements.** Presented in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2025).* [3]

3. Roman Jacome, Leon Suarez, **Romario Gualdrón-Hurtado**, Luis Gonzalez, and Henry Arguello. **Learning to Reconstruct Signals With Inexact Sensing Operator via Knowledge Distillation.** Presented in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2025).* [4]

4. **Romario Gualdrón-Hurtado**, Roman Jacome, Sergio Urrea, Henry Arguello, and Luis Gonzalez. **Learning Point Spread Function Invertibility Assessment for Image Deconvolution.** Presented in *32nd European Signal Processing Conference (EUSIPCO 2024).* [5]

5. **Romario Gualdrón-Hurtado**, Henry Arguello, and Jorge Bacca. **Deep Learned Non-Linear Propagation Model for Compressive Spectral Imaging.** Presented (not indexed) in *LatinX in Computer Vision Research Workshop (LXCV) at IEEE International Conference on Computer Vision (ICCV 2023).*

---

[3] Emmanuel Martinez et al. “Compressive Imaging Reconstruction via Conditional Diffusion Model With Augmented Measurements”. In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2025, pp. 1–5. DOI: 10.1109/ICASSP49660.2025.10889114.

[4] Roman Jacome et al. “Learning to Reconstruct Signals With Inexact Sensing Operator via Knowledge Distillation”. In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2025, pp. 1–5. DOI: 10.1109/ICASSP49660.2025.10887652.

[5] Romario Gualdrón-Hurtado et al. “Learning Point Spread Function Invertibility Assessment for Image Deconvolution”. In: *2024 32nd European Signal Processing Conference (EUSIPCO)*. 2024, pp. 501–505. DOI: 10.23919/EUSIPCO63174.2024.10715342.

6. Emmanuel Martinez, Roman Jacome, **Romario Gualdrón-Hurtado**, Iñaki Esnaola, and Henry Arguello. **Compressive Sensing with Augmented Measurements via Generative Self-Distillation.** Accepted in *IEEE Statistical Signal Processing Workshop (SSP 2025).*

7. Henry Arguello, Roman Jacome*, **Romario Gualdrón-Hurtado***, and Leon Suarez. **NPN: Non-Linear Projections of the Null-Space for Inverse Problems.** Submitted to *The Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025).* *Equal contribution.

8. **Romario Gualdrón-Hurtado***, Roman Jacome*, Leon Suarez, and Henry Arguello. **DIPA: Distilled Preconditioned Algorithms for Solving Imaging Inverse Problems.** Submitted to *The Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025).* *Equal contribution.

9. **Romario Gualdrón-Hurtado**, Roman Jacome, Leon Suarez, and Henry Arguello. **Deep Distillation Gradient Preconditioning for Inverse Problems.** Soon to be submitted to *10th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP 2025).*

INTRODUCTION

Inverse problems are central to a wide range of applications in fields such as medical imaging, computational imaging, and computer vision [6,7]. These problems involve reconstructing or estimating an original signal from measurements that are typically low-dimensional and noisy, which is a fundamental challenge in many domains [8]. These problems often make direct acquisition of the original signal difficult due to existing technological constraints.

One of the main challenges in addressing inverse problems is their ill-posed nature. An ill-posed problem is one where the solution does not necessarily exist, is not unique, or does not depend continuously on the input data. In the context of inverse problems, this often translates to situations where the measurements are insufficient to guarantee a suitable reconstruction of the original signal, particularly when the measurement dimension is much smaller than the signal of interest [9]. This discrepancy can lead to significant errors, especially in the presence of noise.

Traditionally, solving these problems has heavily relied on optimization algorithms that use gradient descent and regularization techniques that incorporate prior knowledge of the signal. Although these approaches are effective, they often require the

---

[6] Andrés Jerez, Miguel Márquez, and Henry Arguello. "Adaptive coded aperture design for compressive computed tomography". In: *Journal of Computational and Applied Mathematics* 384 (2021), p. 113174.

[7] Sergio Urrea et al. "Optical Solutions for Spectral Imaging Inverse Problems with a Shift-Variant System." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 4157–4164.

[8] Curtis R Vogel. *Computational methods for inverse problems*. SIAM, 2002.

[9] David F Shanno and Kang Hoh Phua. "Matrix conditioning and nonlinear optimization". In: *Mathematical Programming* 14 (1978), pp. 149–160.

capture of many low-dimensional measurements to achieve satisfactory results [10]. However, they also face significant limitations, including dependence on potentially inaccurate prior information, and practical constraints on data acquisition, whether physical or temporal [11].

To bridge the gap between traditional methods and the necessity of more advanced solutions for inverse problems, preconditioning emerges as a critical area of focus. Preconditioning techniques aim to modify the problem before solving it, improving the ill-conditioning of the fidelity term. The fidelity term is an ℓ_2-norm that measures the difference between the low-dimensional measurement and propagation of the original signal, often modeled with an ill-conditioned sensing matrix [12]. The preconditioning helps to stabilize the reconstruction step, particularly when dealing with ill-conditioned problems that are inherently sensitive to input variations. Nevertheless, it has been shown that traditional linear preconditioning methods may not be sufficient for the complex and noisy data typically encountered in applications such as medical imaging or computer vision [13,14]. This leads to the exploration of more sophisticated, non-linear approaches that can more effectively handle the dynamics of real-world data, setting the stage for introducing advanced preconditioning methods

---

[10] Rob Kettler. "Analysis and comparison of relaxation schemes in robust multigrid and preconditioned conjugate gradient methods". In: *Multigrid Methods: Proceedings of the Conference Held at Köln-Porz, November 23–27, 1981*. Springer. 2006, pp. 502–534.

[11] Arvind Narayanan and Vitaly Shmatikov. "Robust de-anonymization of large sparse datasets". In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE. 2008, pp. 111–125.

[12] Michele Benzi. "Preconditioning techniques for large linear systems: a survey". In: *Journal of computational Physics* 182.2 (2002), pp. 418–477.

[13] Andrew Godfrey. "Steps toward a robust preconditioning". In: *32nd Aerospace Sciences Meeting and Exhibit*. 1994, p. 520.

[14] Martin J Gander. "On the origins of linear and non-linear preconditioning". In: *Domain decomposition methods in science and engineering XXIII*. Springer. 2017, pp. 153–161.

such as the non-linear gradient preconditioning discussed in this research.

Based on these challenges, there is a clear gap in the research on more adaptable and robust solutions for inverse problems. This work introduces a method that leverages the capabilities of deep learning to enhance the preconditioning process, moving beyond linear methods to offer a solution that can handle the complexities of data more effectively.

This work began with the formulation of a methodology called Measurement Augmentation, which consists of generating complementary measurements to the real ones, which when included in the fidelity term of a PnP-ADMM optimization algorithm behaved as a nonlinear preconditioning method of the gradient. As a consequence, a second more direct nonlinear preconditioning methodology was explored, performing the validation with the FISTA reconstruction algorithm, where by means of a preconditioning operator it was possible to improve the convergence and the quality of the reconstruction. The inclusion of the knowledge distillation technique allowed an improvement beyond that obtained using an end-to-end approach because the teacher can guide in a way closer to the student. The preconditioning operator can be linear or non-linear, and both approaches were tested for optimality. It was compared with different state-of-the-art preconditioning methods, linear, data-driven, and nonlinear. The inverse problems that were validated are super-resolution, single-pixel camera, and magnetic resonance imaging.

In summary, this work advances the field of inverse problems by developing and validating a deep learning-based distillation algorithm that can overcome the limitations posed by ill-conditioned sensing matrices and enhance signal recovery performance. This work fills a research gap in computational imaging and sets the stage for more reliable and efficient solutions.

OBJECTIVES

General objective: To design a nonlinear gradient preconditioning method for inverse problems with a deep learning-based distillation algorithm.

Specific objectives

- To mathematically model an inverse problem including its acquisition process and state-of-the-art recovery algorithms.
- To formulate a nonlinear gradient preconditioning method that employs knowledge distillation theory for the selected inverse problem.
- To implement the designed nonlinear gradient preconditioning method to improve the convergence and performance of state-of-the-art recovery algorithms.
- To validate the designed nonlinear gradient preconditioning method and compare with traditional preconditioning methods.

1. THEORETICAL BACKGROUND

1.1. Inverse Problems in Imaging

Inverse problems are relevant in many areas of science and engineering, serving as the backbone for a variety of applications in computational imaging, such as super-resolution, single-pixel camera, and magnetic resonance imaging [15,16,17,18], as shown in Fig. 1. These problems typically involve recovering an unknown signal $\mathbf{x} \in \mathbb{R}^n$ from observed measurements $\mathbf{y} \in \mathbb{R}^m$ (with $m \ll n$), which can be indirect, incomplete, and noisy, as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \epsilon, \tag{1}$$

where $\mathbf{H} \in \mathbb{R}^{m \times n}$ is the sensing matrix and $\epsilon \in \mathbb{R}^m$ is the sensor noise. The main challenge in inverse problems is their ill-posedness; solutions don't exist for all data, are often non-unique, or are unstable. This instability means small perturbations in input data can lead to large variations in output results, difficulting the reconstruction step [9].

To recover an estimate $\hat{\mathbf{x}}$ of the original image $\mathbf{x}$ from the measurements $\mathbf{y}$, the following optimization problem is formulated

---

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

[16] Marco F Duarte et al. "Single-pixel imaging via compressive sampling". In: *IEEE signal processing magazine* 25.2 (2008), pp. 83–91.

[17] Marinus T Vlaardingerbroek and Jacques A Boer. *Magnetic resonance imaging: theory and practice*. Springer Science & Business Media, 2013.

[18] W Clem Karl et al. "The Foundations of Computational Imaging: A signal processing perspective". In: *IEEE Signal Processing Magazine* 40.5 (2023), pp. 40–53.

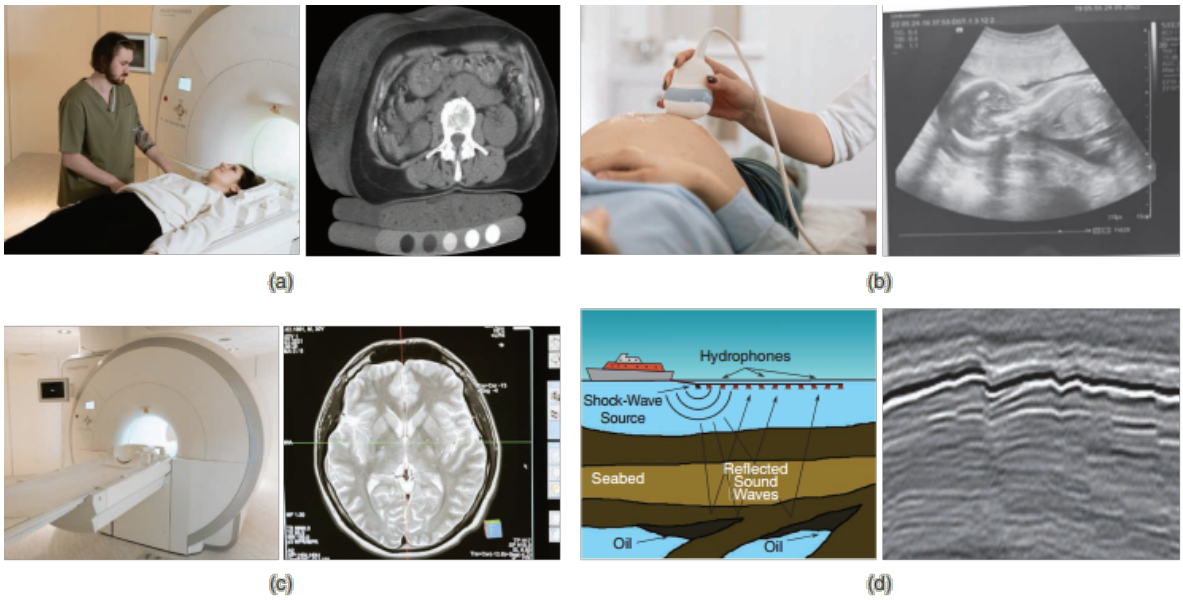


Figure 1. Examples of inverse problems in imaging. (a) X-ray tomography. (b) Ultrasound imaging. (c) Magnetic resonance imaging. (d) Seismic imaging. Adapted from [18].

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} g(\mathbf{x}) + \lambda h(\mathbf{x}), \tag{2}$$

where $g(\mathbf{x})$ is the data fidelity term, ensuring consistency with the measurements, usually $g(\mathbf{x}) = \frac{1}{2}\|\mathbf{H}\mathbf{x} - \mathbf{y}\|_2^2$ and $h(\mathbf{x})$ is a regularizer which constrains the set of feasible solutions by imposing prior knowledge or assumptions (Sec. 1.1.2) about the solution to stabilize the inversion process and ensure the uniqueness and stability of the solution [19]. The regularization parameter $\lambda > 0$ controls the regularization strength. The main challenge is to solve inverse problems efficiently and accurately, and several methods have been proposed (Sec. 1.1.1).

---

[19] Mario Bertero, Patrizia Boccacci, and Christine De Mol. *Introduction to inverse problems in imaging*. CRC press, 2021.

1.1.1. Convex Optimization-Based Methods In many imaging inverse problems, the solution is obtained by minimizing an objective function of the form in (2). When both the data fidelity term $g(\mathbf{x})$ and the regularizer $h(\mathbf{x})$ are convex, a variety of gradient descent–based methods can be applied to efficiently solve the problem. Below, three commonly used approaches are described.

Gradient Descent Methods

- **Steepest Gradient Descent:**

  Steepest Gradient Descent is a straightforward iterative method that updates the solution along the negative gradient direction of the objective function. When minimizing

  $$f(\mathbf{x}) = g(\mathbf{x}) + \lambda h(\mathbf{x}),$$

  the update rule is given by

  $$\mathbf{x}^k = \mathbf{x}^{k-1} - \alpha \nabla f(\mathbf{x}^{k-1}), \tag{3}$$

  where $\alpha > 0$ is the step size. Note that this approach requires both $g(\mathbf{x})$ and $h(\mathbf{x})$ to be differentiable, which may limit its applicability when $h(\mathbf{x})$ is non-smooth.

- **Fast Iterative Shrinkage-Thresholding Algorithm (FISTA):**

  FISTA is an accelerated variant of the standard gradient descent method designed for composite objectives, where $g(\mathbf{x})$ is smooth and $h(\mathbf{x})$ is convex but possibly non-smooth [20]. FISTA introduces a momentum term to speed up con-

[20] Amir Beck and Marc Teboulle. “A fast iterative shrinkage-thresholding algorithm for linear inverse problems”. In: *SIAM journal on imaging sciences* 2.1 (2009), pp. 183–202.

vergence. The iterations are given by

$$\begin{aligned} \mathbf{y}^k &= \mathbf{x}^{k-1} + \frac{t_{k-1}-1}{t_k}(\mathbf{x}^{k-1} - \mathbf{x}^{k-2}), \\ \mathbf{x}^k &= \mathrm{prox}_{\alpha\lambda h}\Big(\mathbf{y}^k - \alpha\nabla g(\mathbf{y}^k)\Big), \end{aligned} \tag{4}$$

where the sequence $\{t_k\}$ is updated as

$$t_k = \frac{1+\sqrt{1+4t_{k-1}^2}}{2},$$

α is the step size, and $\mathrm{prox}_{\alpha\lambda h}$ denotes the proximal operator associated with h.

- **Alternating Direction Method of Multipliers (ADMM):**
  ADMM tackles the optimization problem by splitting the objective into subproblems that are easier to solve. By introducing an auxiliary variable $\mathbf{z}$, the problem in (2) is reformulated as

$$\begin{aligned} \min_{\mathbf{x},\mathbf{z}} \quad & g(\mathbf{x}) + \lambda h(\mathbf{z}), \\ \text{s.t.} \quad & \mathbf{x} = \mathbf{z}. \end{aligned} \tag{5}$$

  The corresponding augmented Lagrangian is

$$\mathcal{L}_\rho(\mathbf{x},\mathbf{z},\mathbf{u}) = g(\mathbf{x}) + \lambda h(\mathbf{z}) + \frac{\rho}{2}\|\mathbf{x}-\mathbf{z}+\mathbf{u}\|_2^2, \tag{6}$$

  where $\mathbf{u}$ is the scaled dual variable and $\rho > 0$ is a penalty parameter. ADMM proceeds by alternating updates for $\mathbf{x}$, $\mathbf{z}$, and $\mathbf{u}$ until convergence, making it particularly effective when the subproblems associated with $g(\mathbf{x})$ and $h(\mathbf{x})$ can be solved efficiently [21].

[21] Stephen Boyd et al. “Distributed optimization and statistical learning via the alternating direction

FISTA was chosen for its balance between convergence speed and ease of integration with modern regularization techniques in imaging inverse problems. In addition, its compatibility with Plug-and-Play approaches makes it ideal for modern image restoration models [20].

1.1.2. Regularization Techniques (Priors) In inverse problems as formulated in (2), the choice of the regularizer $h(\mathbf{x})$ is critical to obtain stable and meaningful solutions. Common regularization techniques include those that promote sparsity or smoothness in the recovered signal. Below is a brief description of several widely used regularizers along with their mathematical formulations:

- **L1-norm (Basis Pursuit):** The ℓ_1-norm promotes sparsity by penalizing the absolute values of the coefficients, effectively driving many of them to zero. It is defined as

$$h(\mathbf{x}) = \|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|, \tag{7}$$

  which is effective in obtaining sparse representations [22].

- **Tikhonov (L2):** Also known as ridge regression, Tikhonov regularization enforces smoothness by penalizing the ℓ_2-norm (or energy) of the solution:

$$h(\mathbf{x}) = \|\mathbf{x}\|_2^2 = \sum_{i=1}^{n} x_i^2. \tag{8}$$

  This regularizer stabilizes the inversion process, mitigating issues related to

---

method of multipliers". In: *Foundations and Trends® in Machine learning* 3.1 (2011), pp. 1–122.

[22] Mark Schmidt. "Least squares optimization with L1-norm regularization". In: *CS542B Project Report* 504.2005 (2005), pp. 195–221.

ill-posedness [23].

- **Total Variation (TV):** The TV regularizer penalizes large gradients in the image, thus preserving important structures such as edges while smoothing out noise [24]. In its anisotropic form, it is defined as

$$\mathrm{TV}(\mathbf{x}) = \sum_{i,j} \left(|x_{i+1,j} - x_{i,j}| + |x_{i,j+1} - x_{i,j}| \right), \tag{9}$$

  and in its isotropic form as

$$\mathrm{TV}(\mathbf{x}) = \sum_{i,j} \sqrt{(x_{i+1,j} - x_{i,j})^2 + (x_{i,j+1} - x_{i,j})^2}. \tag{10}$$

- **Wavelet-based Regularization:** This approach exploits the sparsity of images in the wavelet domain. Let $\mathbf{W}$ denote the wavelet transform operator; then the regularizer is given by

$$h(\mathbf{x}) = \|\mathbf{W}\mathbf{x}\|_1 = \sum_{i=1}^{n} |(\mathbf{W}\mathbf{x})_i| \,. \tag{11}$$

  By promoting sparsity in the wavelet coefficients, this method effectively captures localized and multiscale image features.

The non-smooth nature of these regularizers poses challenges when solving (2) directly. To address this, variable splitting methods like ADMM [21], or FISTA [20], are employed. These techniques decouple the data fidelity term $g(\mathbf{x})$ from the regularization term $h(\mathbf{x})$ by introducing an auxiliary variable (e.g., $\mathbf{z}$), and then use proximal

---

[23] Gene H Golub, Per Christian Hansen, and Dianne P O'Leary. "Tikhonov regularization and total least squares". In: *SIAM journal on matrix analysis and applications* 21.1 (1999), pp. 185–194.

[24] David Strong and Tony Chan. "Edge-preserving and scale-dependent properties of total variation regularization". In: *Inverse problems* 19.6 (2003), S165.

operators to efficiently handle the non-smooth components. Moreover, regularizers such as Tikhonov not only impose smoothness but also integrate prior knowledge about the expected characteristics of the solution, thereby mitigating issues of non-uniqueness and instability [25,26].

1.1.3. Hybrid Optimization-Regularization Methods Hybrid methods bridge the gap between traditional model-based optimization and modern denoising techniques by integrating the strengths of both approaches. These methods leverage state-of-the-art denoisers as implicit priors to handle non-smooth regularization terms, thereby improving the robustness and accuracy of the recovered solutions in inverse problems.

- **Plug-and-Play Optimization (PnP)[27]:** PnP algorithms provide a flexible framework for solving computational imaging inverse problems by integrating model-based inversion methods with denoising algorithms as priors. PnP methods take a different approach by substituting the proximal operator of the regularizer with a general-purpose denoiser D_σ, enabling the use of advanced denoising techniques, such as BM3D [28] or neural-network-based denoisers [29], as implicit

---

[25] Martin Benning and Martin Burger. "Modern regularization methods for inverse problems". In: *Acta numerica* 27 (2018), pp. 1–111.

[26] Bangti Jin, Peter Maaß, and Otmar Scherzer. "Sparsity regularization in inverse problems". In: *Inverse Problems* 33.6 (2017).

[27] Singanallur V. Venkatakrishnan, Charles A. Bouman, and Brendt Wohlberg. "Plug-and-Play priors for model based reconstruction". In: *2013 IEEE Global Conference on Signal and Information Processing*. 2013, pp. 945–948. DOI: 10.1109/GlobalSIP.2013.6737048.

[28] Kostadin Dabov et al. "Image denoising with block-matching and 3D filtering". In: *Image processing: algorithms and systems, neural networks, and machine learning*. Vol. 6064. SPIE. 2006, pp. 354–365.

[29] Harold C Burger, Christian J Schuler, and Stefan Harmeling. "Image denoising: Can plain neural

priors. In the case that $g(\mathbf{x}) = \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2$, the proximal step is replaced by

$$\mathbf{x}^k = \mathrm{D}_\sigma(\mathbf{x}^{k-1} - \alpha\mathbf{H}^\top(\mathbf{H}\mathbf{x}^{k-1} - \mathbf{y})) \tag{12}$$

where $\alpha > 0$ is the stepsize.

- **Regularization by Denoising (RED)** [30]**:** RED also leverages image denoisers to address imaging inverse problems. RED introduces a smoothness-based regularization term $h(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top(\mathbf{x} - \mathrm{D}_\sigma(\mathbf{x}))$, defined by the inner product of the image and its denoising residual. The gradient of $h(\mathbf{x})$ is approximated as the residual $\nabla h(\mathbf{x}) = \mathbf{x} - \mathrm{D}_\sigma(\mathbf{x})$. In contrast to PnP methods, which are limited in their choice of optimization techniques, RED offers more freedom in selecting an optimization method for solving the inverse problem. For instance, using a gradient descent step, the update for the RED objective is given by

$$\mathbf{x}^k = \mathbf{x}^{k-1} - \alpha(\mathbf{H}^\top(\mathbf{H}\mathbf{x}^{k-1} - \mathbf{y}) - \lambda(\mathbf{x}^{k-1} - \mathrm{D}_\sigma(\mathbf{x}^{k-1}))). \tag{13}$$

1.1.4. Deep Learning approach Deep learning has revolutionized the way to solve inverse problems by providing powerful data-driven techniques that can handle the complexity and high dimensionality typically associated with these problems [31,32]. Inverse problems in fields like medical imaging, seismic exploration, and as-

---

networks compete with BM3D?". In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 2392–2399.

[30] Yaniv Romano, Michael Elad, and Peyman Milanfar. "The little engine that could: Regularization by denoising (RED)". in: *SIAM Journal on Imaging Sciences* 10.4 (2017), pp. 1804–1844.

[31] Gregory Ongie et al. "Deep learning techniques for inverse problems in imaging". In: *IEEE Journal on Selected Areas in Information Theory* 1.1 (2020), pp. 39–56.

[32] Hemant K Aggarwal, Merry P Mani, and Mathews Jacob. "MoDL: Model-based deep learning architecture for inverse problems". In: *IEEE transactions on medical imaging* 38.2 (2018), pp. 394–

trophysics often involve reconstructing signals or images from incomplete or noisy measurements, which are inherently challenging due to their ill-posed nature [33]. The application of convolutional neural networks, recurrent neural networks, and other deep learning architectures has shown significant promise in these areas [34]. These networks are capable of learning complex patterns in data, providing robust feature extraction and significant improvements over traditional methods that often require hand-crafted features and extensive domain expertise [1].

One of the key advantages of using deep learning for inverse problems is the ability to directly learn the mapping from the measured data to the desired output, thereby eliminating the need for explicit inversion formulas. A deep neural network is composed of interconnected layers, where each layer applies a linear transformation followed by a non-linear activation. This can be represented as

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \sigma_L\left(\mathbf{W}_L\,\sigma_{L-1}\left(\mathbf{W}_{L-1}\left(\cdots\sigma_1\left(\mathbf{W}_1\mathbf{x}\right)\cdots\right)\right)\right),$$

with $\boldsymbol{\theta} = \{\mathbf{W}_1, \ldots, \mathbf{W}_L\}$ denoting the set of trainable weights and σ_ℓ the non-linear function applied at layer ℓ. The network is trained on a dataset $\{(\mathbf{y}_i, \mathbf{x}_i)\}_{i=1}^N$, where $\mathbf{y}_i$ represents the input and $\mathbf{x}_i$ the corresponding output, by minimizing a loss function that typically takes the form of an ℓ_2-norm error:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \|\mathcal{N}_{\boldsymbol{\theta}}(\mathbf{y}_i) - \mathbf{x}_i\|_2^2 .$$

Through the backpropagation algorithm, the error is propagated backward across

---

405.

[33] Gunther Uhlmann. “Inverse problems: seeing the unseen”. In: *Bulletin of Mathematical Sciences* 4 (2014), pp. 209–279.

[34] Kevin Gurney. *An introduction to neural networks*. CRC press, 2018.

the layers, enabling an iterative update of the weights to achieve minimal error. Deep learning models, particularly those involving autoencoders and generative adversarial networks [35,36], have been effectively used to learn such mappings, often outperforming classical methods in terms of accuracy and efficiency [37]. Furthermore, the emergence of deep learning has facilitated the development of new methodologies in handling inverse problems, such as unsupervised and semi-supervised learning approaches [38,39]. These methods are particularly useful when labeled data is scarce or expensive to obtain [40,41,42]. Overall, deep learning offers a robust framework for addressing several challenges presented by inverse problems, pushing the boundaries in computational imaging.

---

[35] Pierre Baldi. "Autoencoders, unsupervised learning, and deep architectures". In: *Proceedings of ICML workshop on unsupervised and transfer learning*. JMLR Workshop and Conference Proceedings. 2012, pp. 37–49.

[36] Ian Goodfellow et al. "Generative adversarial networks". In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[37] Bo Zhu et al. "Image reconstruction by domain-transform manifold learning". In: *Nature* 555.7697 (2018), pp. 487–492.

[38] Julián Tachella, Dongdong Chen, and Mike Davies. "Unsupervised learning from incomplete measurements for inverse problems". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 4983–4995.

[39] Julián Tachella, Dongdong Chen, and Mike Davies. "Sensing theorems for unsupervised learning in linear inverse problems". In: *Journal of Machine Learning Research* 24.39 (2023), pp. 1–45.

[40] Chunwei Tian et al. "Deep learning on image denoising: An overview". In: *Neural Networks* 131 (2020), pp. 251–275.

[41] Zhihao Wang, Jian Chen, and Steven CH Hoi. "Deep learning for image super-resolution: A survey". In: *IEEE transactions on pattern analysis and machine intelligence* 43.10 (2020), pp. 3365–3387.

[42] Po-Yu Liu and Edmund Y Lam. "Image reconstruction using deep learning". In: *arXiv preprint arXiv:1809.10410* (2018).

1.2. Preconditioning Techniques

Preconditioning techniques are essential for improving the numerical stability and efficiency of the algorithms used to solve inverse problems, particularly when dealing with ill-posed systems [43]. These methods modify the system of equations to enhance the condition number of the problem matrix, thereby accelerating the convergence of iterative methods used for solving these equations. Preconditioning is crucial in scenarios where direct methods fail due to computational infeasibility or where iterative methods converge too slowly due to poor conditioning of the sensing matrix. As illustrated in Figure 2, preconditioning reshapes the optimization space so that fewer steps are required for convergence, compared to the original ill-conditioned problem [44,45].

Condition number

A key concept in assessing the stability of an inverse problem is the *condition number*. The condition number of a matrix provides a quantitative measure of how sensitive the solution is to small perturbations in the input data [46]. For a given matrix $\mathbf{A}$, the condition number is defined as

$$\kappa(\mathbf{A}) = \frac{\sigma_{\max}(\mathbf{A})}{\sigma_{\min}(\mathbf{A})},$$

---

[43] Bjørn Fredrik Nielsen and Kent-Andre Mardal. "Efficient preconditioners for optimality systems arising in connection with inverse problems". In: *SIAM Journal on Control and Optimization* 48.8 (2010), pp. 5143–5177.

[44] Andrew Charles Jones. *Conjugate Gradients*. `https://andrewcharlesjones.github.io/journal/conjugate-gradients.html`. [Accessed: 17-Mar-2025]. 2023.

[45] Youngdo Lee. *[DL] Natural Gradient*. `https://leeyngdo.github.io/blog/deep-learning/2024-02-01-natural-gradient`. [Accessed: 17-Mar-2025]. 2024.

[46] Alan K Cline et al. "An estimate for the condition number of a matrix". In: *SIAM Journal on Numerical Analysis* 16.2 (1979), pp. 368–375.

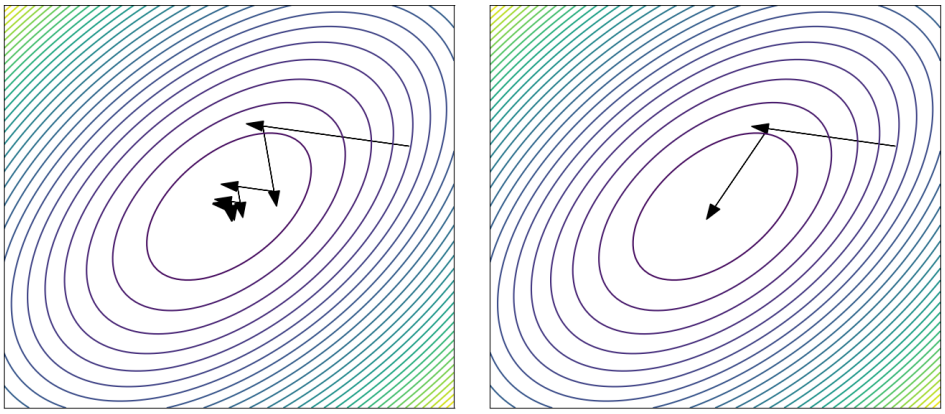


Figure 2. Visualization of an optimization problem in the loss-function space, comparing an ill-conditioned system (left) and a preconditioned system (right). By improving the conditioning of the system, preconditioning reduces the number of iterations needed for gradient-based methods to converge. Adapted from [44,45].

where $\sigma_{\max}(\mathbf{A})$ and $\sigma_{\min}(\mathbf{A})$ denote the largest and smallest singular values of $\mathbf{A}$, respectively. These singular values are obtained from the Singular Value Decomposition (SVD) of $\mathbf{A}$ [47]:

$$\mathbf{A} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^{\top},$$

where:

- $\mathbf{U} \in \mathbb{R}^{n\times n}$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$ are orthogonal matrices.
- $\mathbf{S} = \mathrm{diag}(s_1, s_2, \ldots, s_n)$ is a diagonal matrix whose entries $s_1 \geq s_2 \geq \cdots \geq s_n \geq 0$ are the singular values of $\mathbf{A}$.

A high condition number indicates that the matrix is ill-conditioned, meaning that even small errors or noise in the measurements can result in significant deviations in the reconstructed solution. By applying effective preconditioning strategies, one can reduce the condition number of the system matrix, thereby improving both the

[47] Gilbert W Stewart. "On the early history of the singular value decomposition". In: *SIAM review* 35.4 (1993), pp. 551–566.

convergence rate of iterative solvers and the overall quality of the reconstruction.

1.2.1. Gradient Preconditioning Building upon the foundation of general preconditioning techniques, gradient preconditioning specifically optimizes the convergence properties of gradient-based optimization methods. This advanced technique enhances the efficiency of gradient descent by adjusting the gradient computation to better suit the specific structure of the problem [48].

Gradient preconditioning adjusts the rate at which different directions in the parameter space are updated during the optimization process, addressing the ill-conditioning often observed in high-dimensional optimization problems. This adjustment helps prevent certain directions in the gradient space from dominating others, leading to more balanced updates and faster convergence [49]. The gradient preconditioning is achieved by applying the preconditioning linear operator to the gradient of Eq. (1) as follows

$$\hat{\mathbf{x}}^k = \hat{\mathbf{x}}^{k-1} - \alpha \mathbf{P}\Big(\mathbf{H}^\top\big(\mathbf{H}\hat{\mathbf{x}}^{k-1} - \mathbf{y}\big)\Big). \tag{14}$$

In the context of deep learning, gradient preconditioning involves techniques like adaptive learning rate adjustments and sophisticated second-order methods such as Kronecker-factored Approximate Curvature, which dynamically scale the gradient updates based on the inverse approximations of the Hessian matrix, thereby speed-

---

[48] Elena Caraba. “Preconditioned conjugate gradient algorithm”. In: (2008).

[49] Xi-Lin Li. “Preconditioned stochastic gradient descent”. In: *IEEE transactions on neural networks and learning systems* 29.5 (2017), pp. 1454–1466.

ing up the training process [50,51]. For instance, the gradient preconditioning of (12) and (13) is given by

$$\mathbf{x}^k = \mathrm{D}_\sigma(\mathbf{x}^{k-1} - \alpha\mathbf{P}(\mathbf{H}^\top(\mathbf{H}\mathbf{x}^{k-1} - \mathbf{y}))), \tag{15}$$

and

$$\mathbf{x}^k = \mathbf{x}^{k-1} - \alpha\mathbf{P}(\mathbf{H}^\top(\mathbf{H}\mathbf{x}^{k-1} - \mathbf{y})) - \lambda(\mathbf{x}^{k-1} - \mathrm{D}_\sigma(\mathbf{x}^{k-1})) \tag{16}$$

respectively.

1.2.1.1. Linear Gradient Preconditioning Linear preconditioning methods, such as incomplete Cholesky decomposition or Jacobi methods, have traditionally been employed to simplify the problem matrix while maintaining a balance between computational overhead and convergence rate improvement [52,53]. These methods are effective for symmetric positive definite matrices but often fall short in more complex scenarios, such as those involving non-symmetric or indefinite matrices [54]. A general case of the preconditioned inverse problem is defined as

$$\min_{\mathbf{x}} \|\mathbf{P}^{\frac{1}{2}}(\mathbf{H}\mathbf{x} - \mathbf{y})\|_2^2 + \lambda h(\mathbf{x}), \tag{17}$$

---

[50] Nikolaos Tselepidis, Jonas Kohler, and Antonio Orvieto. "Two-level K-FAC preconditioning for deep learning". In: *arXiv preprint arXiv:2011.00573* (2020).

[51] W Hu et al. "Preconditioned non-linear conjugate gradient method for frequency domain full-waveform seismic inversion". In: *Geophysical Prospecting* 59.3 (2011), pp. 477–491.

[52] Heinz Rutishauser. "The Jacobi method for real symmetric matrices". In: *Numerische Mathematik* 9.1 (1966), pp. 1–10.

[53] David S Kershaw. "The incomplete Cholesky—conjugate gradient method for the iterative solution of systems of linear equations". In: *Journal of computational physics* 26.1 (1978), pp. 43–65.

[54] Tomer Garber and Tom Tirer. "Image Restoration by Denoising Diffusion Models with Iteratively Preconditioned Guidance". In: *arXiv preprint arXiv:2312.16519* (2023).

where $\mathbf{P}$ is the linear preconditioning operator.

Hessian Matrix Preconditioning [55,56]: This approach is based on traditional Newton or quasi-Newton methods [57] in optimization. Here the preconditioner is based on the second derivative of the cost function i.e. $\nabla^2_{\mathbf{x}} g(\mathbf{x}) = \mathbf{H}^\top \mathbf{H}$. The preconditioner is defined as $\mathbf{P} = (\mathbf{H}^\top \mathbf{H})^{-1}$.

As the complexity of data in inverse problems has increased, particularly with the emergence of large-scale imaging systems and high-dimensional data sets, the need for more sophisticated preconditioning techniques has become apparent. Nonlinear preconditioning methods, which adapt the preconditioner based on the properties of the problem at hand, have shown promise in addressing these challenges. These methods are designed to be more flexible and robust, handling variations in the data or the underlying model more effectively [58].

1.2.1.2. Nonlinear Gradient Preconditioning Nonlinear preconditioning represents a significant advance in the field of numerical optimization. It offers a flexible and powerful tool for enhancing the solution of complex inverse problems by dynamically adapting preconditioning strategies based on the current state or iteration of the

---

[55] Ioannis Dassios, Kimon Fountoulakis, and Jacek Gondzio. "A preconditioner for a primal-dual newton conjugate gradient method for compressed sensing problems". In: *SIAM Journal on Scientific Computing* 37.6 (2015), A2783–A2812.

[56] Jeffrey A Fessler and Scott D Booth. "Conjugate-gradient preconditioning methods for shift-variant PET image reconstruction". In: *IEEE transactions on image processing* 8.5 (1999), pp. 688–699.

[57] Eldad Haber, Uri M Ascher, and Doug Oldenburg. "On optimization techniques for solving nonlinear inverse problems". In: *Inverse problems* 16.5 (2000), p. 1263.

[58] Lulu Liu, David E Keyes, and Rolf Krause. "A note on adaptive nonlinear preconditioning techniques". In: *SIAM Journal on Scientific Computing* 40.2 (2018), A1171–A1186.

problem-solving process [59]. This approach transforms the optimization landscape in a way that improves the convergence characteristics of iterative solvers, making it particularly effective for practical applications where models are highly sensitive to input data variations. Nonlinear preconditioning is adaptable to various types of data and models, and it is especially valuable in distributed computing environments where it helps manage computational resources more effectively [14].

The continual evolution of preconditioning techniques, from basic linear methods to sophisticated nonlinear and adaptive approaches, plays a critical role in the ongoing development of computational methods for inverse problems. These techniques not only enhance the efficiency and robustness of numerical algorithms but also expand their applicability across diverse scientific challenges [11].

Gradient preconditioning and nonlinear preconditioning techniques have addressed the critical challenge of convergence in numerical optimization, a central element in the iterative methods used for inverse problems. By improving the conditioning of the problem or the optimization landscape, these advanced preconditioning methods have significantly reduced the computational overhead and improved the robustness of solutions against variations in data quality and model specifications [49,59]. The nonlinear gradient preconditioning can be implemented by incorporating a neural network $\mathcal{N}_{\hat{\theta}}(\cdot)$ to replace the linear operator $\mathbf{P}$ in Eq. (14), thereby enhancing adaptability as shown below:

$$\hat{\mathbf{x}}^k = \hat{\mathbf{x}}^{k-1} - \alpha \mathcal{N}_{\hat{\theta}}\Big(\mathbf{H}^\top\big(\mathbf{H}\hat{\mathbf{x}}^{k-1} - \mathbf{y}\big)\Big). \tag{18}$$

---

[59] Feng-Nan Hwang and Xiao-Chuan Cai. “Improving robustness and parallel scalability of Newton method through nonlinear preconditioning”. In: *Domain decomposition methods in science and engineering*. Springer, 2005, pp. 201–208.

Polynomial Preconditioning [60,61,62]: This preconditioner is designed such that $\|\mathbf{I} - \mathbf{P}\mathbf{H}^\top\mathbf{H}\|_2^2$, which implies that the PM will minimize the eigenvalues of $\mathbf{I} - \mathbf{P}\mathbf{H}^\top\mathbf{H}$ which accelerates the convergence of the algorithm. Mainly, here $\mathbf{P} = p(\mathbf{H}^\top\mathbf{H})$, where the operator $p(\mathbf{T}) = \sum_{i=1}^{n} \varrho(\sigma_i)\mathbf{u}_i\mathbf{v}_i^\top$, where $\mathbf{u}_i, \mathbf{v}_i$ are the singular vectors respectively and ϱ is polynomial function of a degree q. Some works have proposed different ways to choose the coefficients of the polynomial function. One of the most common choices is the Chebyshev polynomials.

1.3. Knowledge Distillation

Knowledge distillation is a technique used to transfer knowledge from a complex model (teacher) to a simpler, more efficient model (student)[63], as shown in Fig. 3. This approach is increasingly relevant in inverse problems where deploying large deep learning models may be computationally prohibitive or inefficient, or when fast inference is required for continuous decision making, such as in medical imaging [63,64,65].

---

[60] Siddharth S Iyer et al. "Polynomial preconditioners for regularized linear inverse problems". In: *SIAM Journal on Imaging Sciences* 17.1 (2024), pp. 116–146.

[61] Hong Ye Tan et al. "Provably convergent plug-and-play quasi-Newton methods". In: *SIAM Journal on Imaging Sciences* 17.2 (2024), pp. 785–819.

[62] Olin G Johnson, Charles A Micchelli, and George Paul. "Polynomial preconditioners for conjugate gradient calculations". In: *SIAM Journal on Numerical Analysis* 20.2 (1983), pp. 362–376.

[63] Jianping Gou et al. "Knowledge distillation: A survey". In: *International Journal of Computer Vision* 129.6 (2021), pp. 1789–1819.

[64] Dian Qin et al. "Efficient medical image segmentation based on knowledge distillation". In: *IEEE Transactions on Medical Imaging* 40.12 (2021), pp. 3820–3831.

[65] Morghan Hartmann, Hasan Farooq, and Ali Imran. "Distilled deep learning based classification of abnormal heartbeat using ECG data through a low cost edge device". In: *2019 IEEE symposium on computers and communications (ISCC)*. IEEE. 2019, pp. 1068–1071.

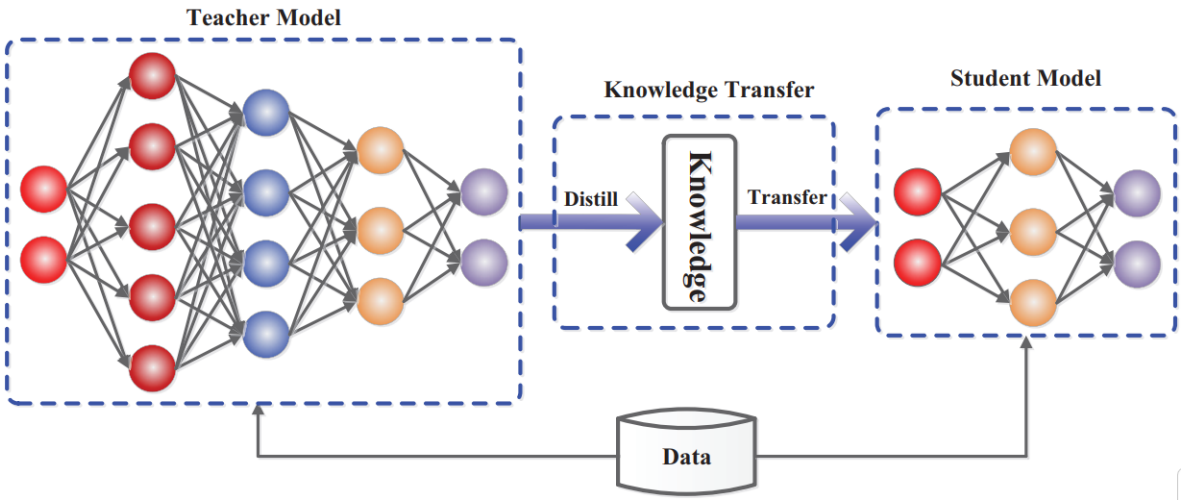


Figure 3. The generic teacher-student framework for knowledge distillation. Adapted from [63].

By distilling the knowledge of a large model into a smaller one, comparable accuracy can be achieved while substantially reducing computational cost and memory requirements [66].

Knowledge distillation is an area of research with significant implications for the field of inverse problems [67]. As these techniques advance, they not only facilitate the deployment of powerful models in constrained environments but also enhance the ability of these models to handle complex inverse problems efficiently and effectively [68].

Knowledge Distillation can be used to learn a preconditioning operator from a less

---

66 Samuel Stanton et al. "Does knowledge distillation really work?" In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 6906–6919.

67 Leon Suarez-Rodriguez, Roman Jacome, and Henry Arguello. "Highly Constrained Coded Aperture Imaging Systems Design Via a Knowledge Distillation Approach". In: *arXiv e-prints* (2024), arXiv–2406.

68 Jathushan Rajasegaran et al. "Self-supervised knowledge distillation for few-shot learning". In: *arXiv preprint arXiv:2006.09785* (2020).

restrictive virtual scenario (teacher) which improves gradient conditioning. In this case the methodology can be reinterpreted, as was done in [67,4,69], where the main difference between the teacher and the student is not the number of parameters or the complexity of the model, but the amount of information you have about the inverse problem you want to solve.

1.4. Quantitative Metrics

This work presents the quantitative reconstruction results using the peak-signal-to-noise ratio (PSNR) metric, which has a maximum value up to infinity, allowing us to make an accurate comparison between reconstructions.

Peak-signal-to-noise ratio (PSNR): To compute the PSNR between the reconstructed image, $\hat{\mathbf{x}}$, and the ground truth image, $\mathbf{x}$, the following equation is used:

$$\mathsf{PSNR}(\mathbf{x}, \hat{\mathbf{x}}) = 10 \cdot \log_{10} \left(\frac{\mathsf{MAX}_{\mathbf{x}}^2}{\mathsf{MSE}(\mathbf{x}, \hat{\mathbf{x}})} \right), \tag{19}$$

where the maximum possible pixel value is set as $\mathsf{MAX}_{\mathbf{x}} = 255$ (for 8-bit images) and the mean squared error (MSE) is defined by

$$\mathsf{MSE}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{n^2} \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2. \tag{20}$$

A PSNR value of ∞ dB indicates identical images, and in practice, values above 30 dB generally indicate that the differences are hardly perceptible to the human eye.

---

[69] Leon Suarez-Rodriguez, Roman Jacome, and Henry Arguello. “Distilling Knowledge for Designing Computational Imaging Systems”. In: *arXiv preprint arXiv:2501.17898* (2025).

2. Gradient Preconditioning via Measurement Augmentation

Initially, the Measurement Augmentation [2] methodology was proposed to improve the conditioning of the inverse problem by generating complementary measurements to the originally acquired real ones. Then, these synthetic measurements are concatenated with the real ones to achieve a better reconstruction. This approach is related to Knowledge Distillation by the inclusion of a virtual teacher neural network that improves the quality of the reconstruction of the real student by generating synthesized measurements. The measurement augmentation technique is depicted in Fig. 4.

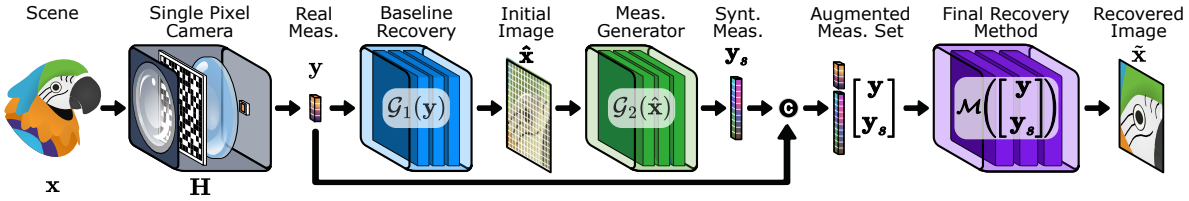


Figure 4. ***Measurement augmentation technique.*** The **scene** $\mathbf{x}$ is propagated by the **single pixel camera** $\mathbf{H}$ to obtain the **real measurements** $\mathbf{y}$, which are then used as input of the **baseline recovery** $\mathcal{G}_1(\cdot)$ that gives an **initial image** $\hat{\mathbf{x}} = \mathcal{G}_1(\mathbf{y})$. This initial estimation is then passed to the **measurement generator** $\mathcal{G}_2(\cdot)$ to obtain the **synthetic measurements** $\mathbf{y}_s$. Both measurement sets are concatenated to form an **augmented measurement set**, $\mathbf{y}_a = \left[\mathbf{y}^\top, \mathbf{y}_s^\top\right]^\top$, that can be employed in any **final recovery method** $\mathcal{M}(\mathbf{y}_a)$ to obtain a better **recovered image** $\tilde{\mathbf{x}}$. Author's own figure, published in [2].

2.1. Measurement Augmentation Technique

This method can be applied to any compressive sensing system and recovery method, but this work focuses on the single-pixel camera (SPC) [70]. To define the sensing process for SPC, let $\mathbf{x}$ represent the vectorized image. The rows of the sensing matrix consist of vectorized coded apertures, denoted as $\mathbf{h}_i \in \mathbb{R}^n, i = 1, \ldots, m$, leading to

---

[70] Graham M Gibson, Steven D Johnson, and Miles J Padgett. "Single-pixel imaging 12 years on: a review". In: *Optics express* 28.19 (2020), pp. 28190–28208.

the matrix $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_{i-th}, \ldots, \mathbf{h}_m]^\top$, where m represents the number of real measurements. Although the SPC typically uses binary values, this work explores real-valued matrices, such as Gaussian and DCT, for a deeper analysis. Our key insight is to improve the conditioning of the data fidelity term by augmenting the number of rows in the sensing matrix $\mathbf{H}$. The primary objective is to predict $\mathbf{y}_s \in \mathbb{R}^d$ such that $\mathbf{y}_s = \mathbf{S}\mathbf{x}$, where $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_{j-th}, \ldots, \mathbf{s}_d]^\top \in \mathbb{R}^{d \times n}$ represents the synthetic sensing matrix for the augmentation that is incoherent with $\mathbf{H}$, i.e., $\mu(\mathbf{H}, \mathbf{S}) = \max_{i,j} |\langle \mathbf{h}_i, \mathbf{s}_j \rangle|$ is small. Note that in practice, only $\mathbf{H}$ is used for measurement acquisition. Here, synthetic measurements refer to those available only in simulations. Thus, the data fidelity term takes the form:

$$f(\tilde{\mathbf{x}}) = \left\| \begin{bmatrix} \mathbf{y} \\ \mathbf{y}_s \end{bmatrix} - \begin{bmatrix} \mathbf{H} \\ \mathbf{S} \end{bmatrix} \tilde{\mathbf{x}} \right\|_2^2 . \tag{21}$$

However, since direct access to $\mathbf{y}_s$ is not available, the objective is to predict it from the measurements $\mathbf{y}$. To this end, a two-step DNN is employed $\mathcal{G} : \mathbb{R}^m \to \mathbb{R}^d$ such that $\mathcal{G}(\mathbf{y}) = \mathcal{G}_2(\mathcal{G}_1(\mathbf{y})) \approx \mathbf{S}\mathbf{x}$. The DNN $\mathcal{G}$ is trained using the mean squared error (MSE) loss function with the following structure. First, a subnetwork $\mathcal{G}_1^* : \mathbb{R}^m \to \mathbb{R}^n$ is pre-trained to obtain an initial estimation $\hat{\mathbf{x}}$, as follows

$$\mathcal{G}_1^* = \arg\min_{\mathcal{G}_1} \mathbb{E}_{\mathbf{x},\mathbf{y}} \Big[\|\mathbf{x} - \mathcal{G}_1(\mathbf{y})\|_2^2 \Big]. \tag{22}$$

In general, any recovery network can be chosen as the initial method and further improved with our approach. Two additional subnetworks are then optimized: $\mathcal{G}_2 : \mathbb{R}^n \to \mathbb{R}^d$, which estimates the synthetic measurements $\mathbf{y}_s$; and $\mathcal{M} : \mathbb{R}^{m+d} \to \mathbb{R}^n$, which refines the estimation of $\mathbf{x}$ using the concatenated measurements. Both are jointly trained as

$$(\mathcal{G}_2^*, \mathcal{M}^*) = \underset{\mathcal{G}_2,\ \mathcal{M}}{\arg\min}\, \mathbb{E}_{\mathbf{x},\mathbf{y}} \Bigg[\Big|\Big| \mathcal{M}\Bigg(\begin{bmatrix} \mathbf{y} \\ \mathcal{G}_2(\mathcal{G}_1^*(\mathbf{y})) \end{bmatrix} \Bigg) - \mathbf{x} \Big|\Big|_2^2 + \zeta \|\mathcal{G}_2(\mathcal{G}_1^*(\mathbf{y})) - \mathbf{S}\mathbf{x}\|_2^2 \Bigg], \tag{23}$$

where ζ promotes the fidelity of the synthetic measurements and $\mathbf{y}_a = \left[\mathbf{y}^\top, \mathbf{y}_s^\top\right]^\top = \left[\mathbf{y}^\top, \mathcal{G}_2^*(\mathcal{G}_1^*(\mathbf{y}))^\top\right]^\top$ is the *augmented measurement set*. $\mathcal{M}(\cdot)$ can also be a fixed model-based recovery method, as explored in Section 2.2 and 2.3.2. Reprojecting the estimate $\mathbf{x}_0 = \mathcal{G}_1^*(\mathbf{y})$ with the synthetic matrix $\mathbf{S}$ would magnify its errors, which is why the $\mathcal{G}_2$ neural network is used, to reduce the noise of $\mathbf{S}\mathbf{x}_0$ to get closer to $\mathbf{S}\mathbf{x}$. As this stage is performed exclusively in a synthetic way, it doesn't need to be physically implementable (linear and binary propagation operator), but it is necessary to obtain the best possible synthetic measure.

2.2. Augmented (Nonlinear gradient preconditioned) PnP Algorithm

Leveraging the subnetwork, $\mathcal{G}_2^*$, that predicts the augmented measurement set, a preconditioned algorithm is developed based on its gradient. To solve (1), it is preconditioned with pre-trained networks as follows

$$\min_{\tilde{\mathbf{x}}} \overbrace{\left\| \begin{bmatrix} \mathbf{y} \\ \mathcal{G}_2^*(\mathcal{G}_1^*(\mathbf{y})) \end{bmatrix} - \begin{bmatrix} \mathbf{H}\tilde{\mathbf{x}} \\ \mathcal{G}_2^*(\tilde{\mathbf{x}}) \end{bmatrix} \right\|_2^2}^{f_a(\tilde{\mathbf{x}})} + \beta g(\tilde{\mathbf{x}}), \tag{24}$$

where $f_a(\tilde{\mathbf{x}})$ is the augmented fidelity term. The insight behind this structure is that the synthetic measurement generating network $\mathcal{G}_2^*$ was trained under two criteria: to have a good estimation over $\mathbf{y}_s$ and that the concatenation of this synthetic measurement can improve the initial reconstruction $\hat{\mathbf{x}} = \mathcal{G}_1^*(\mathbf{y})$. This approach differs from current approaches in state-of-the-art recovery methods, which aim to develop models to exploit some features over the underlying signal $\mathbf{x}$, while some structure of the low-dimensional measurements is exploited here. The PnP-ADMM method is used

to solve equation (24). To simplify the solution process, the problem is reformulated by introducing an auxiliary variable $\mathbf{x} = \mathbf{v}$ as follows:

$$\min_{\tilde{\mathbf{x}},\mathbf{v}} f_a(\tilde{\mathbf{x}}) + \beta g(\mathbf{v}) + \rho\|\tilde{\mathbf{x}} - \mathbf{v}\|_2^2, \text{ s.t. } \mathbf{v} = \tilde{\mathbf{x}}, \tag{25}$$

where ρ is a penalty parameter. The augmented Lagrangian is

$$L_p(\tilde{\mathbf{x}}, \mathbf{v}, \mathbf{u}) = f_a(\tilde{\mathbf{x}}) + \beta g(\mathbf{v}) + \rho\|\tilde{\mathbf{x}} - \mathbf{v} + \mathbf{u}\|_2^2. \tag{26}$$

where $\mathbf{u} \in \mathbb{R}^n$ is the dual variable. ADMM alternates the optimization between primal $\mathbf{x}, \mathbf{v}$ and dual $\mathbf{u}$ [21], yielding:

$$\begin{aligned}
\tilde{\mathbf{x}}^k &= \arg\min_{\tilde{\mathbf{x}}} L_p(\tilde{\mathbf{x}}, \mathbf{v}^{k-1}, \mathbf{u}^{k-1}), \\
\mathbf{v}^k &= \arg\min_{\mathbf{v}} L_p(\tilde{\mathbf{x}}^k, \mathbf{v}, \mathbf{u}^{k-1}), \\
\mathbf{u}^k &= \mathbf{u}^{k-1} + \tilde{\mathbf{x}}^k - \mathbf{v}^k.
\end{aligned}$$

Given that the $\tilde{\mathbf{x}}$-update does not have a closed-form solution, it is solved using gradient descent. The calculated gradient leads to the Jacobian, $\mathcal{J}_{\tilde{\mathbf{x}}}(\cdot)$, of the pre-trained subnetwork that generates the synthetic measurements, $\mathcal{G}_2^*(\cdot)$, with respect to the reconstruction $\tilde{\mathbf{x}}^{k-1}$. Based on the plug-and-play (PnP) concept, $g(\mathbf{v})$ can be solved for $\mathbf{v}$ without explicitly defining $g(\cdot)$, by replacing it with a denoising algorithm $\mathcal{D}(\cdot)$ leading to a PnP-ADMM formulation [71]. Note that the terms highlighted in blue on Algorithm 1 are the derived terms from using the augmented fidelity term. $\mathcal{J}_{\tilde{\mathbf{x}}}(\mathcal{G}_2^*(\tilde{\mathbf{x}}^{k-1}))$ is calculated through the Automatic differentiation package from PyTorch [72]. This additional term can be interpreted as a **non-linear gradient pre-**

---

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

[72] Adam Paszke et al. "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32 (2019).

Algorithm 1 Augmented PnP-ADMM Algorithm

Require: $\mathbf{H}$, $\mathbf{y}$, α, K, ρ, $\mathcal{G}_1^*$, $\mathcal{G}_2^*$

1: $\tilde{\mathbf{x}}^0 = \mathbf{H}^\top \mathbf{y}$, ▷ Algorithm initialization

2: **for** $k = 1 : K$ **do**

3:
$$\tilde{\mathbf{x}}^k = \tilde{\mathbf{x}}^{k-1} - \alpha \Bigg(- \left[\mathbf{H}^\top, \mathcal{J}_{\tilde{\mathbf{x}}}{}^\top(\mathcal{G}_2^*(\tilde{\mathbf{x}}^{k-1}))\right] \times \left(\begin{bmatrix} \mathbf{y} \\ \mathcal{G}_2^*(\mathcal{G}_1^*(\mathbf{y})) \end{bmatrix} - \begin{bmatrix} \mathbf{H}\tilde{\mathbf{x}}^{k-1} \\ \mathcal{G}_2^*(\tilde{\mathbf{x}}^{k-1}) \end{bmatrix} \right) + \rho(\tilde{\mathbf{x}}^{k-1} - \mathbf{v}^{k-1} + \mathbf{u}^{k-1}) \Bigg)$$

4: $\mathbf{v}^k = \mathcal{D}\left(\tilde{\mathbf{x}}^k - \mathbf{u}^{k-1}\right)$ ▷ Denoising update

5: $\mathbf{u}^k = \mathbf{u}^{k-1} + \tilde{\mathbf{x}}^k - \mathbf{v}^k$ ▷ Dual update

6: **end for**

conditioning which improves the convergence and conditioning of the algorithm [14]. $\mathcal{D}(\cdot)$ is the DnCNN architecture [73] implemented in the DeepInverse library [74]. The step-length is $\alpha = 1 \Big/ \left\| \left[\mathbf{H}^\top, \ \mathcal{J}_{\tilde{\mathbf{x}}}^\top(\mathcal{G}_2^*(\tilde{\mathbf{x}}^{k-1}))\right] \right\|_2$. For this case, the fixed $\mathcal{G}_2^*$ from (23) ensures consistent synthetic measurements between the neural network and model-based approaches.

2.3. Experiments & Results

To validate the effectiveness of the proposed method, the acquisition of an SPC is simulated using Hadamard, Cake-Cutting [75], DCT, and Gaussian $\sim \mathcal{N}(0, \mathbf{I})$ sensing matrices with $n = 1024$. The MNIST dataset [76] is used, with 50000 images for training, 10000 for validation, and testing. In these experiments, $\mathbf{H}$ represents the first m

[73] Kai Zhang et al. "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising". In: *IEEE transactions on image processing* 26.7 (2017), pp. 3142–3155.

[74] Julian Tachella et al. *DeepInverse: A deep learning framework for inverse problems in imaging*. Version latest. June 2023. DOI: 10.5281/zenodo.7982256.

[75] Wen-Kai Yu. "Super sub-Nyquist single-pixel imaging by means of cake-cutting Hadamard basis sort". In: *Sensors* 19.19 (2019), p. 4122.

[76] Li Deng. "The mnist database of handwritten digit images for machine learning research [best of the web]". In: *IEEE signal processing magazine* 29.6 (2012), pp. 141–142.

rows of each matrix, while $\mathbf{S}$ corresponds to the subsequent d rows. Additionally, several ablation studies were conducted to determine the optimal hyperparameter configuration. The DNNs, $\mathcal{G}_1^*(\cdot)$ and $\mathcal{G}_2^*(\cdot)$, are fully connected. $\mathcal{G}_1^*(\cdot)$ has m input units, 2048 hidden units, and n output units. $\mathcal{G}_2^*(\cdot)$ takes the initial estimate from $\mathcal{G}_1^*(\mathbf{y})$ as input, with n input units, 2048 hidden units, and d output units, producing the synthetic measurement $\mathbf{y}_s$. The DNNs were trained for 200 epochs using the Adam optimizer [77] with a learning rate of 5×10^{-4} and a batch size of 750 images. Both neural network and model-based recovery methods can be used as $\mathcal{M}(\cdot)$ to recover the signal from $\mathbf{y}_a$.

2.3.1. Neural network-based recovery A neural network $\mathcal{M}(\cdot)$ is trained using the optimization problem in Eq. (23). Here, a fully connected network with $m + d$ inputs, 2048 hidden units, and n outputs was used. A comparative study of $\zeta = \{0.001, 0.01, 0.1, 1.0, 10, 100\}$ is performed for different compression ratios, defined as m/n and d/n, between the original measurements $\mathbf{y}$ and the synthetic measurements $\mathbf{y}_s$, respectively. The best ζ value for each sensing matrix is reported in Fig. 5, with the baseline at the top of each column. The baseline consists of recovering the signal without additional synthetic measurements. In each column of Fig. 5, the traditional and Cake-Cutting ordering methods for the Hadamard matrix are compared as binary matrices, along with the DCT and Gaussian as real-valued matrices. Based on the best PSNR values (highlighted in **bold**), it can be concluded that for low real compression ratios, a larger number of synthetic measurements, d, is needed to achieve optimal recovery quality compared to scenarios with higher compression ratios. Note that the total compression ratio, $(m + d)/n$, must remain less than or equal to 1.

---

[77] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015.

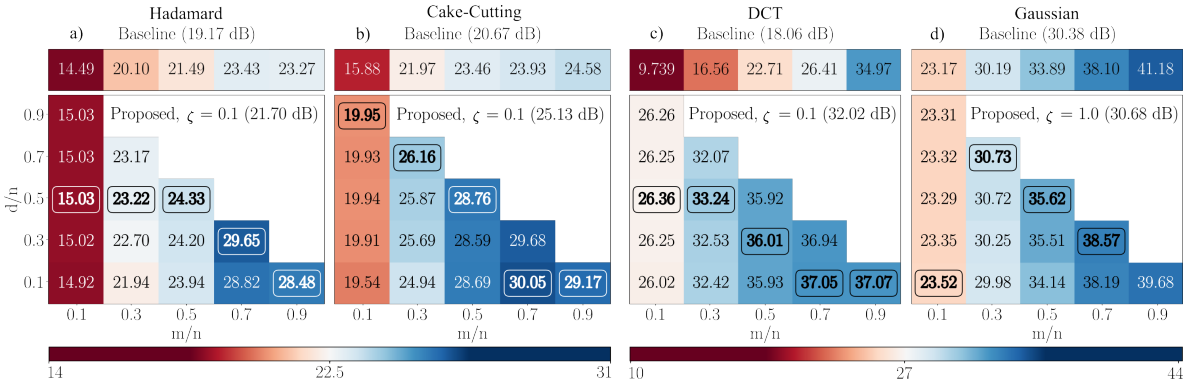


Figure 5. Measurement augmentation performance in terms of peak signal-to-noise ratio (PSNR) using a DNN for signal recovery with binary sensing matrices: a) Hadamard and b) Cake-Cutting ordered Hadamard and real-valued sensing matrices: c) DCT and d) Gaussian. The best recovery PSNR for each ζ experiment is highlighted in **bold** if it surpasses the baseline in the corresponding m/n. In this format, the baseline appears at the top for each sensing matrix, with comparisons made column-wise based on the m/n value. The synthetic compression ratio (d/n) varies by heatmap row. Blank zones indicate cases where $m/n + d/n > 1.0$, exceeding the original signal's information. Average PSNR (dB) for each compression ratio is in parentheses. Author's own figure, published in [2].

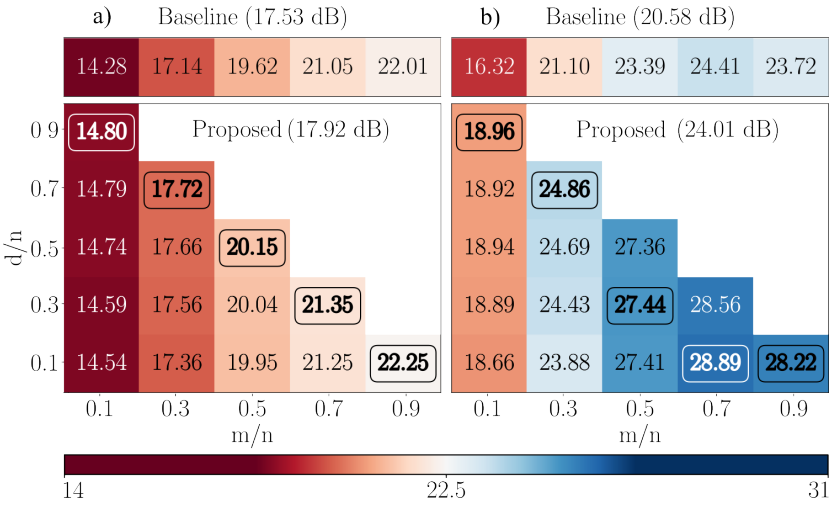


Figure 6. Measurement augmentation robustness with a) 5 and b) 25 dB SNR of Gaussian noise SPC with Cake-Cutting sensing matrix and $\zeta = 1.0$. The proposed method outperforms the baseline in each configuration despite the different noise levels. (See Fig. 5 caption for explanation of this format). Author's own figure, published in [2].

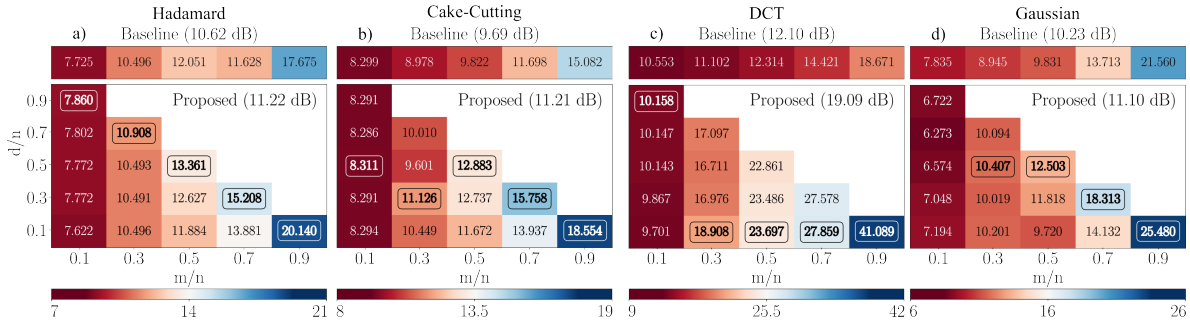


Figure 7. Augmented PnP-ADMM performance in terms of PSNR for signal recovery with binary sensing matrices: a) Hadamard and b) Cake-Cutting ordered Hadamard, and real-valued sensing matrices: c) DCT and d) Gaussian. The best recovery PSNR for a fixed $\zeta = 0.1$ experiment is highlighted on **bold** if it surpasses the baseline in the corresponding compression ratio m/n. (See Fig. 5 caption for explanation of this format). Author's own figure, published in [2].

An analysis of robustness for different noise levels is shown in Fig. 6 with 5 dB and 25 dB signal-to-noise ratio (SNR) white Gaussian noise on the measurements. The model-based noise analysis should behave similarly but with reduced performance compared to the learning-based method in Fig. 6. The proposed method surpasses the baseline in all the compression ratio combinations.

2.3.2. Model-based recovery Here, the proposed approach was validated by using the augmented PnP-ADMM presented in Algorithm 1 as $\mathcal{M}(\cdot)$. In this case, the DNNs $\mathcal{G}_1^*$ and $\mathcal{G}_2^*$ have to be pre-trained and fixed. The gradient of Eq. (24) yields the Jacobian of $\mathcal{G}_2^*$, acting as a non-linear gradient preconditioner for PnP-ADMM. Fig. 7 presents the baseline and results for the different sensing matrices, where the Augmented PnP-ADMM outperforms the measurement-restricted PnP-ADMM in most of the cases. Note that in Fig. 5 as in Fig. 7, the cases where fewer real measurements are acquired, i.e., m/n is low, data-driven, and model-based recovery methods tend to perform better if a larger amount of synthetic measurements, d/n, is added to them, a trend that decreases as the real measurement quantity increases.

2.3.3. Convergence and conditioning analysis The convergence of the different losses in the model-based recovery is presented in Fig. 8. Based on these

results, it can be concluded that when the m/n is low, in this case 0.1, the large null space of the sensing matrix undermines the fidelity term's ability to provide effective guidance. This conclusion is derived from the fact that the recovered signals with measurement-restricted PnP-ADMM (red line) have a lower fidelity loss but a higher ground-truth loss with respect to the Augmented PnP-ADMM.

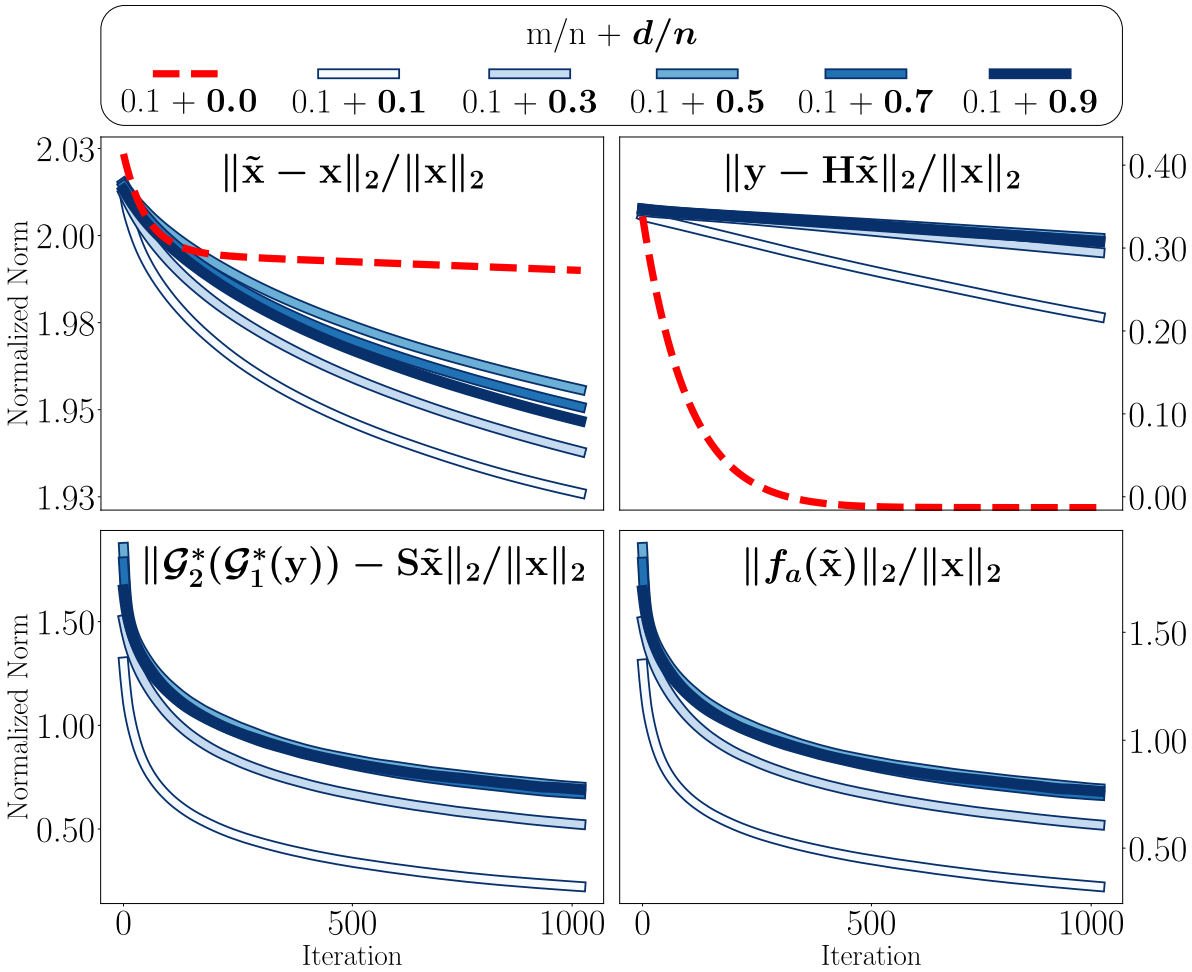


Figure 8. Convergence metrics for Augmented PnP-ADMM performance for the DCT sensing matrix with $m/n = 0.1$. The losses are normalized with respect to the ground-truth signal $\mathbf{x}$. Author's own figure, published in [2].

3. Gradient Preconditioning via Knowledge Distillation

It is possible to design a preconditioning operator (PO), $\mathcal{P}$, using the Knowledge Distillation (KD) methodology. It has proven to be able to outperform other data-driven learning methods such as End-to-End (E2E) [69]. In this case, the same methodology of KD initially proposed by the authors for Deep Learning will not be performed [78], where the main difference between the teacher model and the student model was the number of parameters, which allowed the teacher a better generalization and performance. In this case, a modified approach is proposed for inverse problems, where the main difference between the teacher and the student lies in the number of measurements acquired, the exposure time, or the resolution. That is, the teacher model will be an algorithm with a better-conditioned sensing matrix than the student's, which will allow it to guide the student's learning by PO to its gradients. Before detailing the proposed optimization approach, the teacher and student models will be defined. [79]

3.1. Student Algorithm and Teacher Algorithm setting

Here, the student model of the proposed KD framework is considered to be a gradient-preconditioned student algorithm (GPSA) using a sensing matrix $\mathbf{H}_s \in \mathbb{R}^{m_s \times n}$ which produces a measurement vector $\mathbf{y}_s = \mathbf{H}_s\mathbf{x} + \mathbf{e}_s$. While the teacher model is a teacher algorithm (TA) that recovers the underlying signal $\mathbf{x}$ from a measurement vector obtained with a virtual sensing matrix (infeasible to implement in practice but with high

---

[78] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. *Distilling the Knowledge in a Neural Network*. 2015. arXiv: 1503.02531 [stat.ML].

[79] This methodology was developed in conjunction with PhD student, Roman Jacome, under the supervision of Professor Henry Arguello.

reconstruction performance) $\mathbf{H}_t \in \mathbb{R}^{m_t \times n}$ as $\mathbf{y}_t = \mathbf{H}_t\mathbf{x} + \mathbf{e}_t$. The student model requires the PO since $\mathbf{H}_s$ is more ill-conditioned than the virtual teacher sensing matrix $\mathbf{H}_t$. Here, three inverse problems with different teacher and student sensing matrix settings are presented.

Inverse Problem 1 (Magnetic Resonance Imaging: MRI)**.** *Consider the single-coil MRI, where undersampled k-space measurements are acquired to reduce the acquisition time* [80]*. Here, the teacher and the student are defined as* $\mathbf{H}_s = \mathbf{M}_s\mathbf{F} \in \mathbb{C}^{n\times n}$ *where* $\mathbf{F} \in \mathbb{C}^{n\times n}$ *is the 2D discrete Fourier transform of the signal and* $\mathbf{M}_s \in \mathbb{C}^{n\times n}$ *is a 2D mask sampling the k-space. Similarly, the teacher is* $\mathbf{H}_t = \mathbf{M}_t\mathbf{F} \in \mathbb{C}^{n\times n}$*. The 2D mask matrices* $\mathbf{M}_s$ *and* $\mathbf{M}_t$ *are governed by the so-called acceleration factor (AF) which indicates the subsampled ratio of the masks, i.e.,* $AF_s = \frac{n^2}{\|\mathbf{M}_s\|_0}$ *and* $AF_t = \frac{n^2}{\|\mathbf{M}_t\|_0}$*. The higher the AF, the fewer k-space scans are employed, leading to faster acquisition. Thus, it is considered that* $AF_t \ll AF_s$*, making the teacher infeasible in practice due to the high acquisition time required, but it allows high recovery performance, while the student requires reduced acquisition time.*

Inverse Problem 2 (Single-Pixel Camera: SPC)**.** *The SPC acquires coded projections of the scene using coded apertures* [16]*. Here, the teacher and the student are defined as* $\mathbf{H}_t \in \{-1, 1\}^{m_t\times n}$ *and* $\mathbf{H}_s \in \{-1, 1\}^{m_s\times n}$*, respectively, where* m_t *and* m_s *are the number of snapshots and each row comes from the Hadamard basis. Increasing the number of snapshots causes increased acquisition and processing time. Thus, it is set that* $m_t \gg m_s$ *so that the student can be easily implemented in practice. In this context, the compression ratios are given by* $\gamma_t = \frac{m_t}{n}$ *for the teacher and* $\gamma_s = \frac{m_s}{n}$ *for the student.*

Inverse Problem 3 (Super Resolution: SR)**.** *SR aims to reconstruct a high-resolution*

---

[80] Klaas P Pruessmann et al. “Coil sensitivity encoding for fast MRI”. in: *Proceedings of the ISMRM 6th Annual Meeting, Sydney*. Vol. 1998. 1998.

(HR) image from a low-resolution (LR) observation [15]*. The student and teacher models employ different sensing matrices, where* $\mathbf{H}_s = \mathbf{D}_s\mathbf{B}_s \in \mathbb{R}^{m_s \times n}$ *and* $\mathbf{H}_t = \mathbf{D}_t\mathbf{B}_t \in \mathbb{R}^{m_t \times n}$*. Here,* $\mathbf{D}$ *represents a downsampling operator, and* $\mathbf{B}$ *models the system's blur. The resolution factor (RF) defines the reduction in spatial resolution, given by* $RF_s = \frac{\sqrt{n}}{\sqrt{m_s}}$ *for the student and* $RF_t = \frac{\sqrt{n}}{\sqrt{m_t}}$ *for the teacher. Since* $RF_s \gg RF_t$*, the teacher model is impractical due to its high-resolution input requirement but achieves superior reconstruction, whereas the student operates on feasible LR images.*

Based on these criteria, the PO is designed such that the gradient-preconditioned student algorithm behaves similarly to the virtual teacher algorithm.

3.2. Distilling the preconditioning operator

Our approach is data-driven, where a dataset of clean images $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{I}$ with I images is employed to generate the student and teacher measurements as $\mathcal{Y}_s = \{\mathbf{y}_{s_i} | \mathbf{y}_{s_i} = \mathbf{H}_s\mathbf{x}_i + \mathbf{e}_{s_i}\}_{i=1}^{I}$ and $\mathcal{Y}_t = \{\mathbf{y}_{t_i} | \mathbf{y}_{t_i} = \mathbf{H}_t\mathbf{x}_i + \mathbf{e}_{t_i}\}_{i=1}^{I}$. The optimization of the PO is proposed as

$$\begin{aligned} \mathcal{P}^\star = \underset{\mathcal{P}}{\operatorname{argmin}}\, \mathbb{E}_{\mathbf{x}_i, \mathbf{y}_{s_i}, \mathbf{y}_{t_i}} \mathcal{L}_{KD}\left(\hat{\mathbf{x}}_{\mathcal{P}s_i}, \mathbf{H}_s, \mathbf{H}_t, \hat{\mathbf{x}}_{t_i}, \mathcal{P}\right), \\ \text{s.t } \hat{\mathbf{x}}_{\mathcal{P}s_i} = \texttt{GPSA}\left(\mathcal{P}, \mathbf{H}_s, \mathbf{y}_{s_i}\right) \\ \hat{\mathbf{x}}_{t_i} = \texttt{TA}\left(\mathbf{H}_t, \mathbf{y}_{t_i}\right), \end{aligned} \tag{27}$$

where the notation $\texttt{GPSA}\left(\mathcal{P}, \mathbf{H}_s, \mathbf{y}_{s_i}\right)$ and $\texttt{TA}\left(\mathbf{H}_t, \mathbf{y}_{t_i}\right)$ refers to performing the GPSA and TA respectively. The crucial aspect here is the cost function $\mathcal{L}_{KD}(\cdot)$ that depends on the teacher and student sensing matrices and reconstructions. The proposed cost function is the following

$$\mathcal{L}_{KD} = \mathcal{L}_G(\hat{\mathbf{x}}_{\mathcal{P}s_i}, \mathbf{H}_s, \mathbf{H}_t, \hat{\mathbf{x}}_{t_i}, \mathcal{P}) + \beta\mathcal{L}_I(\hat{\mathbf{x}}_{\mathcal{P}s_i}, \hat{\mathbf{x}}_{t_i}), \tag{28}$$

where $\mathcal{L}_G$ is the gradient loss, and $\mathcal{L}_I$ is the imitation loss.

3.2.1. Gradient loss The gradient loss $\mathcal{L}_G$ aligns the direction of the GPSA's data-fidelity gradient with that of the TA as follows. For matrix inputs, we interpret cosine similarity via vectorization (or equivalently the Frobenius inner product). Denoting $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$, we set

$$\mathcal{S}_c(\mathbf{A}, \mathbf{B}) = \frac{\langle \mathrm{vec}(\mathbf{A}),\, \mathrm{vec}(\mathbf{B}) \rangle}{\|\mathrm{vec}(\mathbf{A})\|_2 \, \|\mathrm{vec}(\mathbf{B})\|_2} = \frac{\mathrm{tr}(\mathbf{A}^\top \mathbf{B})}{\|\mathbf{A}\|_F \, \|\mathbf{B}\|_F}, \tag{29}$$

which yields a scalar in $[-1, 1]$. The loss is then

$$\mathcal{L}_G = \big(1 - \mathcal{S}_c(\, \mathcal{P}\, \nabla g_s(\hat{\mathbf{x}}_{s_i}),\, \nabla g_t(\hat{\mathbf{x}}_{t_i})\,)\big)^2, \tag{30}$$

where $g_s(\hat{\mathbf{x}}_{s_i}) = \|\mathbf{y}_{s_i} - \mathbf{H}_s \hat{\mathbf{x}}_{s_i}\|_2^2$, $g_t(\hat{\mathbf{x}}_{t_i}) = \|\mathbf{y}_{t_i} - \mathbf{H}_t \hat{\mathbf{x}}_{t_i}\|_2^2$ are the student and teacher data fidelity terms. This loss function guides the PO design towards the data fidelity gradient of the preconditioned student to have a similar direction to the well-conditioned teacher gradient.

3.2.2. Imitation loss Finally, the term $\mathcal{L}_I$ is called an imitation loss function since its main objective is that the GPSA obtains the same output of the TA. This term is defined as

$$\mathcal{L}_I = \|\hat{\mathbf{x}}_{\mathcal{P}s_i} - \hat{\mathbf{x}}_{t_i}\|_2^2. \tag{31}$$

Combining these two loss functions allows the PO design to behave like the TA in the GPSA without being highly affected by the physical limitations of the real implementation. The optimization problem is solved using off-the-shelf stochastic gradient

descent algorithms such as Adam [77] or AdamW [81]. Note that unlike traditional PO design, which only aims to improve the algorithm's convergence rate, our approach aims to improve recovery performance by the guidance of the TA with the $\mathcal{L}_I$ loss function and to achieve good convergence rate via the $\mathcal{L}_G$ loss.

3.2.3. Supervised loss The supervised loss, as an alternative to the imitation loss, optimizes the GPSA's PO using the ground truth (GT) label $\mathbf{x}_i$ instead of the teacher's output, defined as

$$\mathcal{L}_S = \|\hat{\mathbf{x}}_{\mathcal{P}s_i} - \mathbf{x}_i\|_2^2. \tag{32}$$

3.2.4. Convergence regularization Although the proposed KD-based PO aims for the student GPSA to imitate the TA dynamics, an improved convergence rate can be ensured by adding regularization functions in (27), inspired by the GPSA and TA convergence analysis. The focus here is on the PnP-FISTA scheme; however, the adaptation of this analysis to the RED-FISTA scheme is similar.

Assumption 1 (Bounded denoiser)**.** *The denoiser* D_σ *is bounded such that for* $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$

$$\|\mathrm{D}_\sigma(\mathbf{x}) - \mathrm{D}_\sigma(\mathbf{z})\|_2 \leq (1+\delta)\|\mathbf{x} - \mathbf{z}\|_2, \tag{33}$$

where $\delta > 0$ *is a numerical constant.*

Usually, the convergence of the PnP algorithm is derived via fixed-point analysis [71]. Here, the convergence of the GPSA with the teacher PnP is analyzed to design a regularization function that minimizes this. First, consider the iterations given by

$$\mathbf{x}_s^k = \mathrm{T}_s(\mathbf{x}_s^{k-1}) = \mathrm{D}_\sigma(\mathbf{x}_s^{k-1} - \alpha\mathcal{P}(\mathbf{H}_s^\top(\mathbf{H}_s\mathbf{x}_s^{k-1} - \mathbf{y}_s))),$$

[81] I Loshchilov. "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101* (2017).

and

$$\mathbf{x}_t^k = \mathrm{T}_t(\mathbf{x}_t^{k-1}) = \mathrm{D}_\sigma(\mathbf{x}_t^{k-1} - \alpha(\mathbf{H}_t^\top(\mathbf{H}_t\mathbf{x}_t^{k-1} - \mathbf{y}_t))).$$

Thus, based on this, a comparison is made to evaluate how the student iterations approximate those of the teacher algorithm.

$$\begin{aligned}
&\left\|\mathbf{x}_s^k - \mathbf{x}_t^k\right\| \\
&= \left\|\mathrm{T}_s\left(\mathbf{x}_s^k\right) - \mathrm{T}_t\left(\mathbf{x}_t^k\right)\right\| \\
&= \left\|\mathrm{D}_\sigma\left(\mathbf{x}_s^k - \alpha\mathcal{P}\left(\mathbf{H}_s^\top\left(\mathbf{H}_s\mathbf{x}_s^k - \overbrace{\mathbf{H}_s\mathbf{x}}^{\mathbf{y}_s}\right)\right)\right) - \mathrm{D}_\sigma\left(\mathbf{x}_t^k - \alpha\left(\mathbf{H}_t^\top\left(\mathbf{H}_t\mathbf{x}_t^k - \underbrace{\mathbf{H}_t\mathbf{x}}_{\mathbf{y}_t}\right)\right)\right)\right\| \\
&\leq (1+\delta)\left\|\left(\mathbf{I} - \mathcal{P}(\mathbf{H}_s^\top\mathbf{H}_s)\right)\mathbf{x}_s^k - \left(\mathbf{I} - \mathbf{H}_t^\top\mathbf{H}_t\right)\mathbf{x}_t^k\right\| + \underbrace{\left\|\left(\mathcal{P}(\mathbf{H}_s^\top\mathbf{H}_s) - \mathbf{H}_t^\top\mathbf{H}_t\right)\mathbf{x}\right\|}_{R_C(\mathcal{P},\mathbf{H}_s,\mathbf{H}_t,\mathbf{x})},
\end{aligned}$$

where $\mathbf{x}$ is the GT image, the third line employs Assumption 1, and the fourth line follows the triangle inequality. The highlighted term is the proposed convergence regularization term. This function is chosen as regularization since it depends only on the GT image available during training and the student and teacher sensing matrices. Thus, with this function, the optimization of the PO is given by

$$\begin{aligned}
\mathcal{P}^\star = \underset{\mathcal{P}}{\operatorname{argmin}}\ &\mathbb{E}_{\mathbf{x}_i,\mathbf{y}_{s_i},\mathbf{y}_{t_i}}\mathcal{L}_{KD}\left(\hat{\mathbf{x}}_{\mathcal{P}s_i},\mathbf{H}_s,\mathbf{H}_t,\hat{\mathbf{x}}_{t_i},\mathcal{P}\right) + \tau R_C(\mathcal{P},\mathbf{H}_s,\mathbf{H}_t,\mathbf{x}) + \mathcal{L}_S(\hat{\mathbf{x}}_{\mathcal{P}s_i},\mathbf{x}_i) \\
\text{s.t}\quad &\hat{\mathbf{x}}_{\mathcal{P}s_i} = \texttt{GPSA}\left(\mathcal{P},\mathbf{H}_s,\mathbf{y}_{s_i}\right) \\
&\hat{\mathbf{x}}_{t_i} = \texttt{TA}\left(\mathbf{H}_t,\mathbf{y}_{t_i}\right),
\end{aligned} \tag{34}$$

where $\tau > 0$ is a regularization parameter.

3.3. Experimental Setup for Inverse Problems

The effectiveness of the proposed approach was validated across three common imaging inverse problems, MRI, super-resolution (SR), and compressed sensing (CS) with the single-pixel camera (SPC), using the RED and PnP algorithm implementations from the DeepInv Library [82] for image recovery. The algorithm used is FISTA [20] due to its balance between convergence speed and ease of integration with modern regularization techniques in inverse problems in imaging. An ablation study was performed for the convergence regularization parameter and then set to $\tau = 1 \times 10^{-3}$ for both RED and PnP experiments. The stepsize for the teacher $\alpha_t = 0.7$ and for the student $\alpha_s = 0.4$ with 20 iterations. The PO was optimized for 50 epochs. The computer used for the simulations has an Intel(R) Xeon(R) W-3223 CPU @ 3.50GHz processor, 48 GB RAM, and an NVIDIA GeForce RTX 3090 with 24 GB VRAM.

Magnetic Resonance Imaging: The FastMRI single-coil knee dataset from [83], preprocessed by [82], was used. This dataset consists of 900 training and 73 testing MRI knee images, each with a size of 320×320. Due to the large size of the PO when using the images at their original resolution, they were resized to 50×50 ($n = 2500$). The AdamW optimizer [81] was employed with a learning rate of 1×10^{-5} and a weight decay of 0.01. The batch size was set to 8 for the RED experiments, while for the PnP experiments, it was set to 32. For all MRI experiments, a 1D Gaussian undersampling mask with binary values was employed.

---

[82] Julian Tachella et al. *DeepInverse: A deep learning framework for inverse problems in imaging*. Version latest. June 2023. DOI: 10.5281/zenodo.7982256.

[83] Florian Knoll et al. "FastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning". en. In: *Radiol. Artif. Intell.* 2.1 (Jan. 2020), e190007.

Single-Pixel Camera: The SPC is used along with the MNIST dataset [76], with $50,000$ images for training and $10,000$ for testing. All images were resized to 32×32. The Adam [77] optimizer was used with a learning rate of 1×10^{-5}. For the RED experiments, the batch size was 50, for the PnP experiments, the batch size was 225. A 2D subsampled Hadamard transform is used as the sensing matrix for all the SPC experiments.

Super-Resolution: Additional experiments were performed for Super-Resolution with the CelebA dataset [84], with $162,770$ images for training and $19,962$ for testing. All images were resized to 110×110. The Adam optimizer [77] was used with a learning rate of 1×10^{-5}. For the RED experiments, the batch size was 3, for the PnP experiments, the batch size was 4.

---

[84] Ziwei Liu et al. "Deep Learning Face Attributes in the Wild". In: *Proceedings of International Conference on Computer Vision (ICCV)*. 2015.

3.4. DIPA: Linearly Distilled Gradient Preconditioned Algorithms

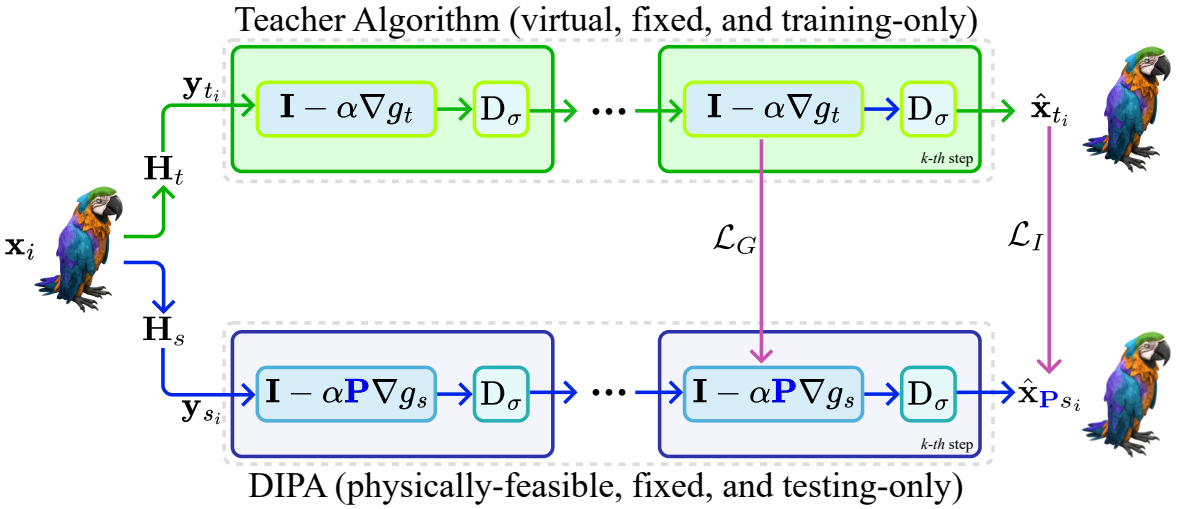


Figure 9. Proposed DIPA framework: A teacher algorithm that employs a well-conditioned sensing matrix, feasible in simulation but impractical for implementation, distills its knowledge to improve the performance of a student algorithm (DIPA), which accounts for a physically implementable sensing matrix, through a preconditioning matrix (PM) $\mathbf{P} \in \mathbb{R}^{n\times n}$. The distillation procedure involves transferring knowledge from the teacher's outputs and the directions of the data fidelity gradients. The baseline student is the case when the PM is $\mathbf{P} = \mathbf{I}$. Author's own figure.

Considering the generality of the distilled preconditioning methodology, it was initially tested with the generation of a preconditioning matrix (PM). In this particular case, the preconditioning operator from Eq. (27) will be formulated as $\mathcal{P} = \mathbf{P} \in \mathbb{R}^{n\times n}$, ensuring compatibility with the gradients of the student algorithm (SA) and TA. This linear formulation facilitates the analysis and observation of what was learned, which is not possible with a neural network. The PM is initialized as an identity matrix $\mathbf{I} \in \mathbb{R}^{n\times n}$, which is also the baseline case, where the SA is not preconditioned. However, this linear approach, although more efficient than the best state-of-the-art methods, scales with the dimension of the images and is $O(n^2)$.

3.4.1. Ablation studies In Fig. 10 an ablation study for different values of acceleration factor (AF) and compression ratio (γ) was performed for both the PnP

Teacher and the DIPA-PnP, where the latter manages to outperform the baseline in most cases and performs best when the (virtual) teacher has as much information as possible. In this figure, the comparisons should be made by columns, and the values enclosed indicate the best reconstruction for a specific student configuration. Combinations with blank zones were not tested because it implied that the student was equally or better conditioned than the teacher.

Fig. 11 show visual reconstructions of the reconstructions and ground-truth in MRI. These images qualitatively and quantitatively reflect the enhancement capability of the DIPA-PnP when guided by the gradient of the well-conditioned PnP Teacher. The sampling mask used for the teacher and the student are shown in the right column. Note that with a reduced amount of information it is possible to improve the baseline by applying a PM learned from the virtual teacher.

An ablation study is performed for the SR task using the CelebA dataset [84]. All images were resized to 110×110. The quantitative results are shown in Table 1, where improvements of up to 14 dB in PSNR are obtained due to the teacher's guidance. Note that in some settings, for instance, $\mathcal{R}_C(\checkmark)$ and $\mathcal{L}_S(\checkmark)$ for $AF_t = 3$ the student outperforms the teacher, this is due in this experiment, the supervised loss uses the ground-truth (GT) during the training which is better guidance than the teacher estimation. However, the GT is not always available.

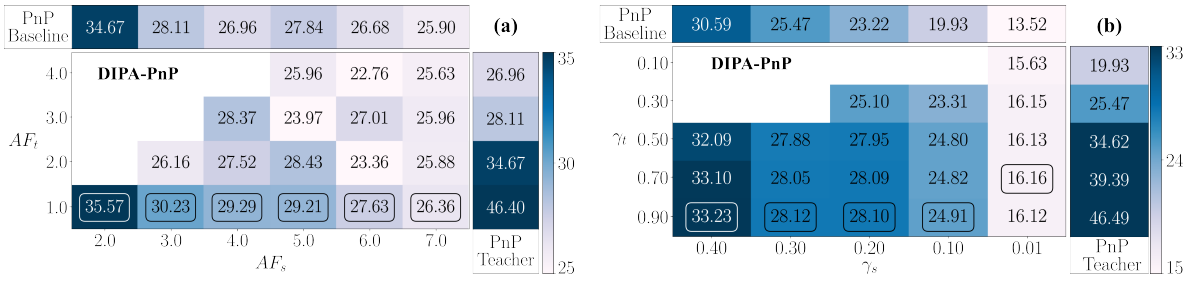


Figure 10. Recovery quality in terms of PSNR for MRI (a) and SPC (b) with different Acceleration Factors (AF) and compression ratios (γ) for Teacher (t) and Student (s). Note that the Baseline PnP-FISTA (top) is outperformed by the DIPA-PnP (center), leveraging the guidance of the PnP Teacher (right). Author's own figure.

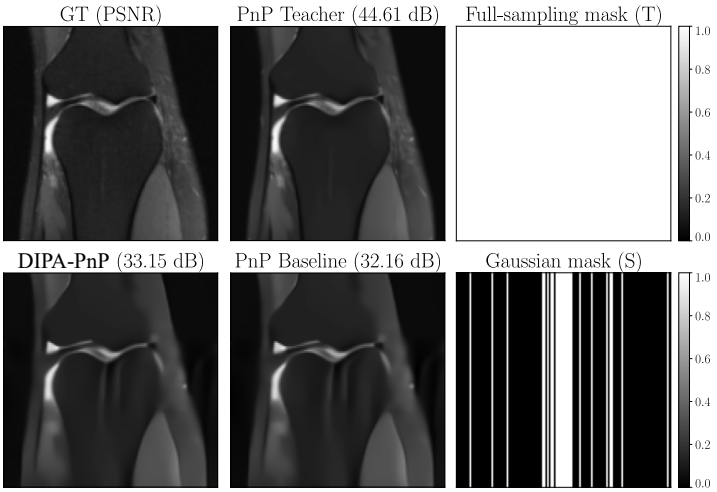


Figure 11. Visual results, PSNR and masks for MRI with a 128×128 spatial resolution with the PnP-FISTA algorithm, $\mathcal{R}_C$(✓), $\mathcal{L}_S$(✗), $AF_s = 5$, and $AF_t = 1$. Author's own figure.

		Super-Resolution					
$\mathcal{R}_C$	$\mathcal{L}_S$	RF_t	**DIPA-PnP**	**Teacher-PnP**	RF_t	**DIPA-RED**	**Teacher-RED**
✓	✓	3	24.90	17.89	3	24.76	17.93
		2	24.95	30.08	2	24.75	30.45
		1	**25.00**	33.97	1	24.81	34.49
✗	✓	3	24.98	17.89	3	24.77	17.93
		2	24.95	30.08	2	24.77	30.72
		1	<u>24.99</u>	33.97	1	<u>24.82</u>	34.49
✓	✗	3	17.56	17.89	3	17.90	18.09
		2	24.51	30.08	2	24.55	30.58
		1	24.92	33.97	1	**25.12**	34.48
✗	✗	3	17.58	17.90	3	17.73	17.93
		2	24.59	30.08	2	24.43	30.72
		1	24.89	33.97	1	24.72	34.49
			Base.			**Base.**	
			11.02			10.83	

Table 1. Ablation study in terms of PSNR for Super-Resolution (110×110) with different RF_t values for the PnP and RED teachers, with $RF_s = 4$ for the students (DIPA-PnP and DIPA-RED) with the CelebA dataset.

Visual results for SR are shown in Fig. 12(a), where the DIPA-RED outperforms the baseline and is considerably closer to the RED Teacher even though its measurement has a resolution 4 times lower than the teacher. It is also possible to use higher-resolution images for SPC. In this study, the BSDS500 dataset was used [85] dataset resized to 128×128 in grayscale. The loss function ablation study is performed. The results in Table 2 show significant improvements by up to 4 dB in PSNR. Some visual results of this experiment are shown in Figure 12(b) for DIPA-RED with $\gamma_s = 0.2$ and $\gamma_t = 0.7$.

$\mathcal{R}_C$	$\mathcal{L}_S$	DIPA-PnP	DIPA-RED
✓	✓	**32.65**	33.45
✗	✓	<u>32.64</u>	**34.99**
✓	✗	32.29	32.82
✗	✗	32.28	<u>33.94</u>
Teacher		38.54	38.73
Baseline		29.33	29.05

Table 2. Ablation study in terms of PSNR for SPC (128×128) with $\gamma_t = 0.7$ for the PnP and RED teachers and $\gamma_s = 0.2$ for the students (DIPA-PnP and DIPA-RED) with the BSDS500 dataset.

3.4.2. State-of-the-art (SOTA) comparison SOTA preconditioning methods were implemented [86,87], ensuring a rigorous and fair comparison. An exhaustive parame-

---

[85] Pablo Arbelaez et al. "Contour detection and hierarchical image segmentation". In: *IEEE transactions on pattern analysis and machine intelligence* 33.5 (2010), pp. 898–916.

[86] Matthias J Ehrhardt, Patrick Fahy, and Mohammad Golbabaee. "Learning preconditioners for inverse problems". In: *arXiv e-prints* (2024), arXiv–2406.

[87] Tomer Garber and Tom Tirer. "Image restoration by denoising diffusion models with iteratively preconditioned guidance". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 25245–25254.

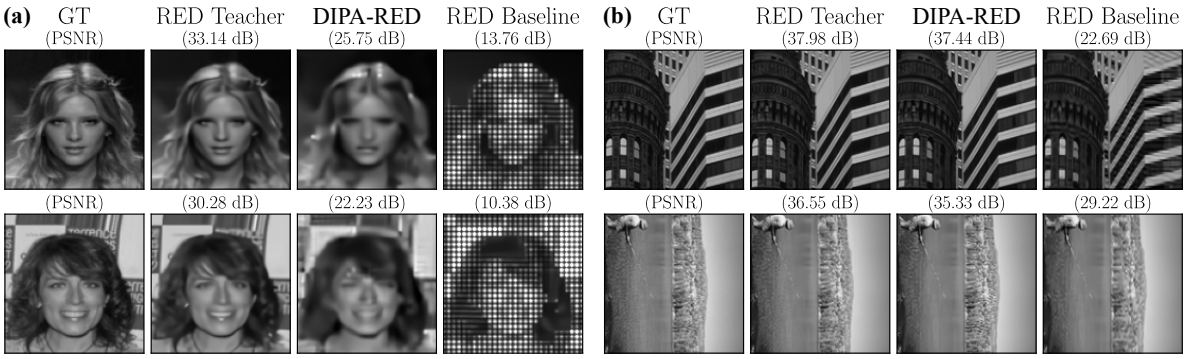


Figure 12. Visual results and PSNR with RED-FISTA for DIPA.
(a) SR (110×110), $RF_t = 1$, $RF_s = 4$, $\mathcal{R}_C$(✓), $\mathcal{L}_S$(✗) with the CelebA dataset.
(b) SPC (128×128), $\gamma_t = 0.7$, $\gamma_s = 0.2$, $\mathcal{R}_C$(✗), $\mathcal{L}_S$(✓) with the BSDS500 dataset.
Author's own figure.

ter search was conducted for each method to optimize performance. The optimization of various preconditioning matrices was incorporated into the reconstruction of SPC measurements at a compression ratio of $\gamma = 0.2$. The methods were trained for 50 epochs, except the Baseline which does not need training. Because of this, the Baseline is the fastest, but with a deficient performance. DIPA-PnP has the highest performance and is the most efficient among the preconditioning methods. 20 iterations of the PnP-FISTA algorithm were run for each method. No method (including DIPA-PnP) uses the convergence regulation or supervised loss proposed in this paper, only those of their authors.

Learned	Method	Formulation	MRI ($AF_s = 5$)	SR ($RF_s = 4$)	SPC ($\gamma_s = 0.2$)
✗	Baseline	$\mathbf{P} = \mathbf{I}_n$	25.77	11.14	22.36
✗	Polynomial	$\mathbf{P} = p(\mathbf{H}^\top \mathbf{H})$	28.07	11.81	23.01
✗	Hessian	$\mathbf{P} = (\mathbf{H}^\top \mathbf{H})^{-1}$	28.16	11.75	22.94
✓	Scalar step	$\mathbf{P_k} = p_k \mathbf{I_n}$	27.88	11.62	23.36
✓	Pointwise	$\mathbf{P_k} = \mathbf{p_k} \odot \mathbf{x}$	27.59	11.19	21.28
✓	Convolutional	$\mathbf{P_k} = \mathbf{p_k} * \mathbf{x}$	28.63	17.55	21.30
✓	Full-linear	$\mathbf{P_k}$	28.15	22.02	26.19
✓	DIPA (Ours)	$\mathbf{P} = \mathbf{P}^\star$	<u>29.21</u>	<u>25.00</u>	<u>32.74</u>

Table 3. State-of-the-art preconditioning comparison for MRI, SR, and SPC, against DIPA.

Some works have designed the PM based solely on the structure of the sensing

matrix $\mathbf{H}$. Here, the proposed PM design is compared with previous approaches. For the experiments involving Polynomial Preconditioning, the implementation from [61] was utilized, employing a polynomial of degree $q = 2$. For the experiments with Hessian Matrix Preconditioning, the pseudoinverse approximation was applied.

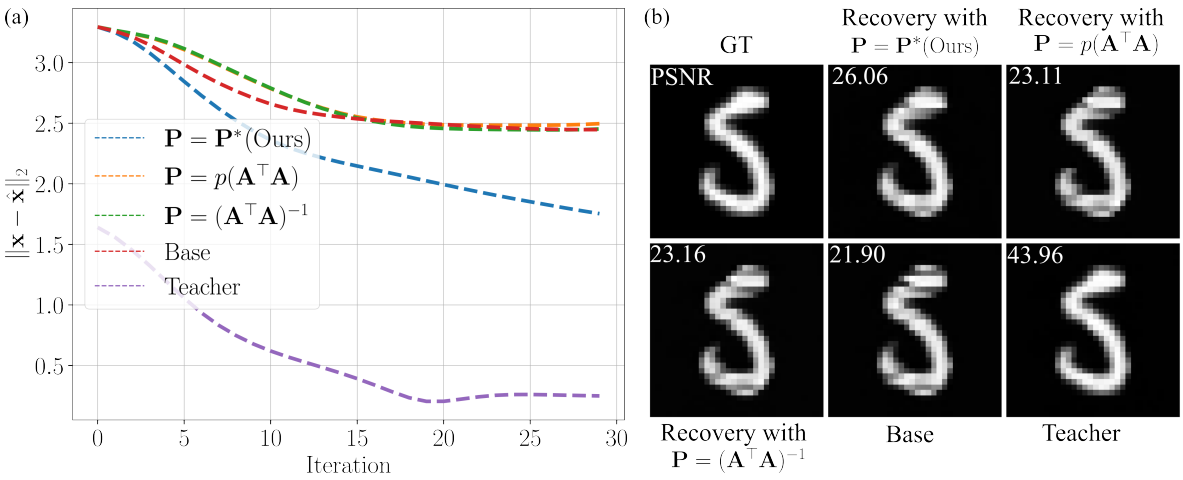


Figure 13. Convergence of the DIPA-PnP method compared with traditional PMs and visual reconstructions for SPC with $\gamma_s = 0.2$. DIPA-PnP used $\mathcal{R}_C(\checkmark)$ and $\mathcal{L}_S(\checkmark)$. Author's own figure.

3.4.2.1. Experiments with multi-coil MRI The proposed method was implemented for multi-coil MRI following similar experimental settings as those used for single-coil MRI, incorporating birdcage-generated coil maps. Table 4 presents the validation results for different acceleration factors and numbers of coil maps. The matrix $\mathbf{H}_t$ was set with $AF_t = 1$ and 15 coil maps. Results show improvements of up to 1 dB.

3.4.3. Convergence and conditioning analysis In Fig. 13, the convergence and visual representations in SPC are shown. A value of $\alpha = 0.1$ was used for the polynomial preconditioning, and $\alpha = 0.05$ for the Hessian matrix. The number of iterations of the GPSA was set to 30. It can be seen that although the traditional PM

Coil maps	$AF_s = 5$		$AF_s = 4$		$AF_s = 3$	
	DIPA-PnP	**Baseline**	**DIPA-PnP**	**Baseline**	**DIPA-PnP**	**Baseline**
5	**32.70**	31.61	**37.33**	36.71	**38.81**	38.32
10	**32.81**	31.78	**37.39**	36.80	**38.87**	38.38
15	**32.83**	31.80	**37.38**	36.80	**38.88**	38.39

Table 4. Performance for multi-coil MRI varying the number of coils with DIPA-PnP optimized with $\mathcal{R}_C(\checkmark)$ and $\mathcal{L}_S(\boldsymbol{\times})$.

methods outperforms the baseline in almost all cases, their improvement is minimal compared with the improvement of the proposed approach. Fig. 14 complements Fig. 13, as they are the same scenario. Analyzing them together, it can be observed that the proposed method (in blue color) although it stagnates in the convergence of its fidelity, like the polynomial preconditioning, the quality of the reconstruction is much better than the latter, since it is taking advantage of the null space in a better way and converging to a better solution. Here is where the improvement of the preconditioning is clearly evidenced, allowing to obtain a better reconstruction from iteration 8, which requires up to 30 iterations for the other methods. The singular values of the different preconditioning methods are varied, and although the baseline and the polynomial seem to have the best condition number, their reconstruction is one of the worst, evidencing a trade-off that is very interesting, since it is not always necessary to lower the condition number to obtain the best reconstruction in the least amount of iterations.

Then, the convergence of the DIPA-PnP method was analyzed in comparison with traditional preconditioning matrices. A compression ratio of $\gamma_s = 0.2$ was used. For the proposed method, the PM was trained using a teacher PnP with $\gamma_t = 0.7$, employing $\mathcal{R}_C(\checkmark)$ and $\mathcal{L}_S(\checkmark)$. In Figure 13, the reconstruction error is plotted at each iteration per method, showing an improved convergence with the proposed DIPA-PnP method. Additionally, a visual reconstruction of the MNIST test set is displayed. The reconstruction shows that the baseline and the traditional preconditioner obtain horizontal artifacts in the image.

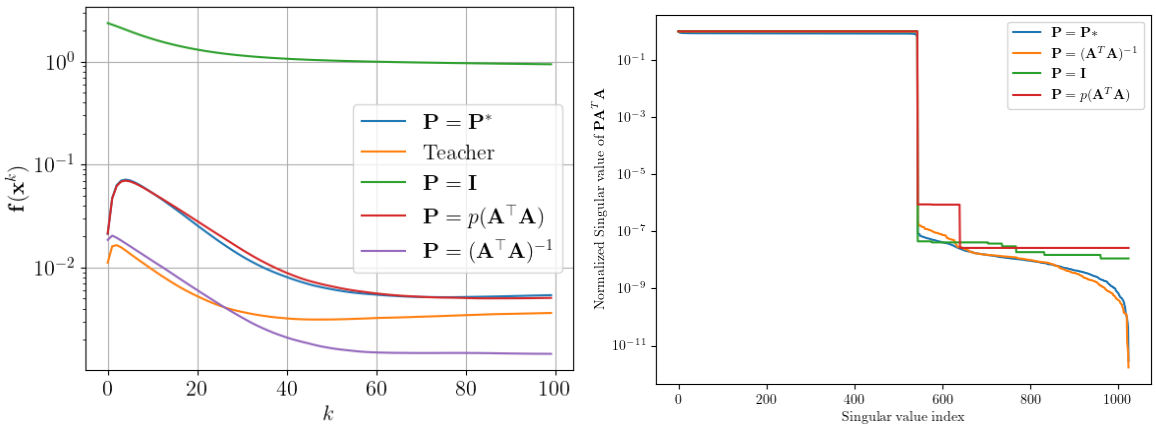


Figure 14. Fidelity term convergence of the DIPA-PnP method compared with traditional PMs and comparison of Singular Values of $\mathbf{PA}^\top\mathbf{A}$, where $\mathbf{A} = \mathbf{H}_s$ for SPC with $\gamma_s = 0.2$. DIPA-PnP used $\mathcal{R}_C(\checkmark)$ and $\mathcal{L}_S(\checkmark)$. Author's own figure.

3.4.4. Visual representation of the PM Fig. 15 shows the logarithmic representation of a 128×128 zoom of the PMs. The resulting SPC, MRI, SR, and higher-resolution MRI PMs are shown. In the left column, the identity matrix is presented for comparison, serving as the baseline with which they were initialized. Note how different structures are generated depending on the imaging task; in SPC square patterns are generated as windows with diagonals. In MRI and higher-resolution MRI, there are areas of varying intensity in the direction of the diagonal, although in the latter it is less present. Finally, in SR vertical areas are generated along the diagonal. Additionally, there is variability in the intensity of the diagonal of each PM. This reflects the ability of the PM to adapt to the imaging task and spatial resolution.

3.4.5. Robustness of the PM The objective is to analyze the robustness of the trained PM when evaluated using a different method than the one it was originally trained with. Particularly, a PM trained with DIPA-RED is tested with both DIPA-RED and DIPA-PnP algorithms. This evaluation is conducted for MRI, using a student with $AF_s = 5$ trained with different teachers with $AF_s = \{1, 2, 3\}$ using DIPA-RED, spatial

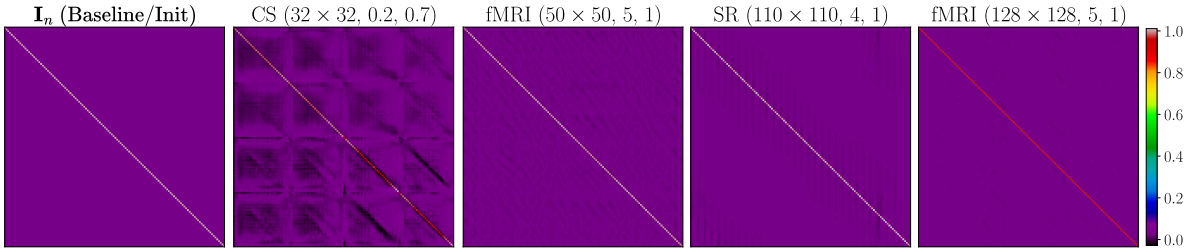


Figure 15. Natural logarithmic representation of the preconditioning matrices (PMs), $log(PM + 1)$, for different tasks and resolutions. Title format: Imaging task (Spatial resolution, Student information, Teacher information). Author's own figure.

Train\Test	AF_t	PnP	RED
RED	1	29.65	27.58
	2	28.05	27.54
	3	29.65	27.31
		Base PnP 27.77	Base RED 27.31

Table 5. Cross validation of the trained PM along different algorithms in MRI.

resolution of 50×50, $\mathcal{R}_C$(✓), $\mathcal{L}_S$(✗). The results are shown in Table 5. The results show that, while the PM was trained with DIPA-RED, using the PM with DIPA-PnP led to a good performance (outperforming the baseline). This allows us to conclude that the PM is suitable for different algorithms.

3.5. D^{2}GP: Deep Distillation for Non-Linear Gradient Preconditioning

In section 3.4, it was observed that the DIPA methodology improved the quality of the reconstruction. However, the number of PM parameters is $O(n^2)$ in the linear case. Thus, the research question arose: How can a deep learning-based distillation algorithm be effectively applied to improve signal recovery through distilled nonlinear gradient preconditioning in inverse problems, and how does it compare with existing linear preconditioning methods, given the challenges posed by ill-conditioned sensing matrices in inverse problems?

The nonlinear preconditioning operator (NPO), now a neural network $\mathcal{P}_\theta(\cdot)$, will be reformulated to learn a set of weights, θ, that are less than those of $\mathbf{P}$ thus allowing better quality reconstruction in a more efficient manner:

$$\begin{aligned}\theta^\star = \underset{\theta}{\operatorname{argmin}}\, \mathbb{E}_{\mathbf{x}_i,\mathbf{y}_{s_i},\mathbf{y}_{t_i}} \mathcal{L}_{KD}\left(\hat{\mathbf{x}}_{\mathcal{P}_\theta s_i}, \mathbf{H}_s, \mathbf{H}_t, \hat{\mathbf{x}}_{t_i}, \mathcal{P}_\theta\right) + \tau R_C(\mathcal{P}_\theta, \mathbf{H}_s, \mathbf{H}_t, \mathbf{x}) + \mathcal{L}_S(\hat{\mathbf{x}}_{\mathcal{P}_\theta s_i}, \mathbf{x}_i),\\ \text{s.t } \hat{\mathbf{x}}_{\mathcal{P}_\theta s_i} = \texttt{GPSA}\left(\mathcal{P}_\theta, \mathbf{H}_s, \mathbf{y}_{s_i}\right)\\ \hat{\mathbf{x}}_{t_i} = \texttt{TA}\left(\mathbf{H}_t.\mathbf{y}_{t_i}\right), \end{aligned} \tag{35}$$

3.5.1. Neural Network ablation studies Different neural networks (NNs) that were used as NPOs are shown. Among them are: Multilayer perceptron [88], con-

[88] Frank Rosenblatt. "The perceptron: a probabilistic model for information storage and organization in the brain." In: *Psychological review* 65.6 (1958), p. 386.

volutional NN (CNN) [89], CNN with attention module [90], UNet [91], MultiScale CNN [92], IndiUNet [93], Vision Transformer [94], and ConvNext [95]. For the ablation studies on Fig. 16(a), modifications were made to reduce the number of parameters of each of the NNs, to have fewer parameters than $\mathbf{P}$, and using the configuration that obtained the best possible performance. Because of this, some networks such as ViT, despite their widely known potential, are not able to generalize the gradient mapping from student to teacher. Regardless of the configuration, and because of the hidden neurons, the MLP will have more even with more parameters than $\mathbf{P}$, it does not perform adequately. ConvNeXt is the one that proves to perform better with a reduced number of parameters, due to its exploitation of convolutional networks.

3.5.2. Nonlinear Preconditioning Operator (NPO) setup In Fig. 16(b), you can see some of the configurations that were tested for the ConvNeXt, where the

---

[89] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[90] Sanghyun Woo et al. "Cbam: Convolutional block attention module". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 3–19.

[91] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241.

[92] Qiangqiang Yuan et al. "A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11.3 (2018), pp. 978–989.

[93] Mauricio Delbracio and Peyman Milanfar. "Inversion by direct iteration: An alternative to denoising diffusion for image restoration". In: *arXiv preprint arXiv:2303.11435* (2023).

[94] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[95] Zhuang Liu et al. "A convnet for the 2020s". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 11976–11986.

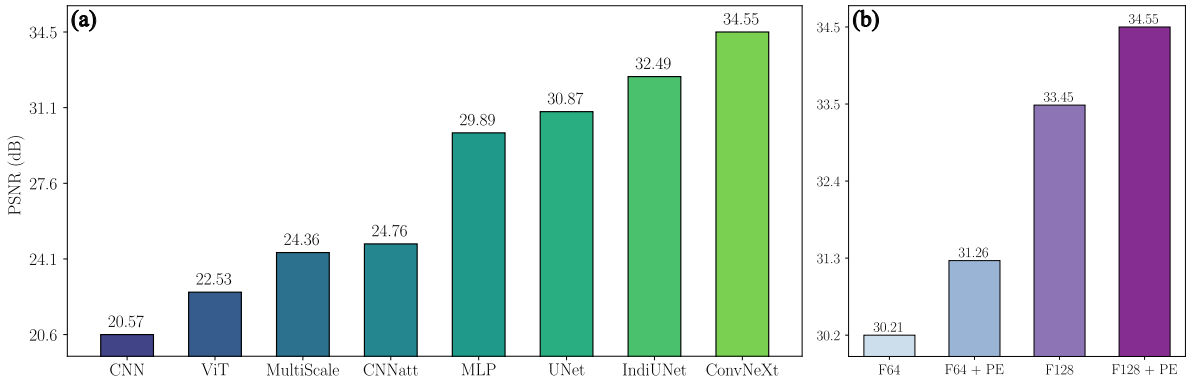


Figure 16. (a) Ablation results in terms of PSNR of different NNs as NPO for SPC. (b) Number of features and positional encoding (PE) usage in ConvNeXt. Author's own figure.

inclusion of the positional encoder (PE), favored the quality of the final reconstructions. The maximum number of features where the efficiency was maintained and the best performance was achieved is 128. Finally, a ConvNeXt network of 5 blocks and 128 features per block was used, with a learning rate of 1×10^{-5}, Adam optimizer, scheduler ReduceLROnPlateau, a residual connection in the last layer, and the PE proposed in IndiUNet [93] so that the network could know and differentiate the gradients according to the current iteration of the algorithm. A representation of the ConvNeXt, adapted from [96,97], is shown in Fig. 17. For MRI, 256 features were required.

3.5.3. Ablation studies of the losses Figs. 18, 19, and 20 show an ablation study of the different losses in SPC, MRI and SR, all using $\mathcal{L}_G$. The combination of losses is specified on the right side of the caption, in the color table. In the legends,

---

[96] Atakan Erdogan. *ConvNeXt: Next Generation of Convolutional Networks*. `https://medium.com/@atakanerdogan305/convnext-next-generation-of-convolutional-networks-325607a08c46`. [Accessed: 17-Mar-2025]. 2021.

[97] Ayesha Kanwal et al. "A hybrid framework for detection of autism using ConvNeXt-T and embedding clusters". In: *The Journal of Supercomputing* 80.6 (2024), pp. 8156–8178.

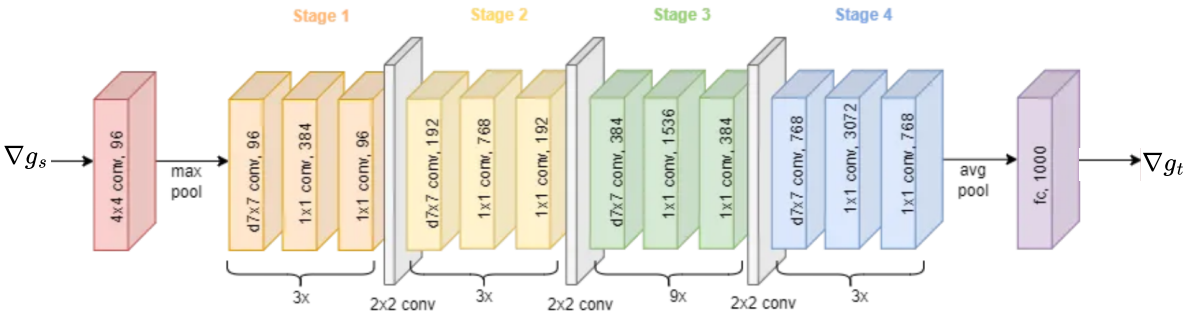


Figure 17. General architecture of the ConvNeXt Neural Network used as NPO. Figure adapted from [96,97]

you can see the two cases proposed as baseline, DIPA with linear preconditioning (green) and E2E with nonlinear preconditioning (red). The PSNR without preconditioning is shown in text at the end of each caption. All figures show a marked improvement over the SOTA methods, the most interesting being the case of SR, in Fig. 20, as there a growth is observed when using the three losses $\mathcal{L}_G$, $\mathcal{L}_I$, and $\mathcal{R}_C$. Fig. 21 shows the visual results in the different tasks, where the proposed method outperforms those of the state-of-the-art, bounded by the Teacher performance.

3.5.4. State-of-the-art comparison An extensive comparison was made of state-of-the-art methods, divided between those that require learning by optimization (DIPA (Sec. 3.4), scalar step[86], pointwise[86], convoutional[86], and full-linear[86]) and those that do not (Baseline ($\mathbf{P} = \mathbf{I}$), Hessian[55,56] and Polynomial[60,61,62]). In addition to the proposed method, its formulation without the use of the Knowledge Distillation methodology is used as baseline, and it is named Nonlinear E2E, since it uses $\mathcal{L}_S$ for its learning, in this case there is no teacher, but the ground-truth is directly accessed during training. In Tables 6, 7, and 8, one can observe the results in terms of PSNR, the number of trainable parameters of each method, and the ratio with respect to the trainable parameters of the NPO of D^{2}GP, for the different SOTA methods, preconditioning the FISTA-PnP algorithm. Specifically, in Table 6 for SPC, it is observed that D^{2}GP is able to outperform the best SOTA method (Full-linear) by more than 8 dB

and with 24 times fewer parameters. Additionally, it outperforms by almost 2 dB the DIPA linear method (proposed by us) with 1.22 times fewer parameters.

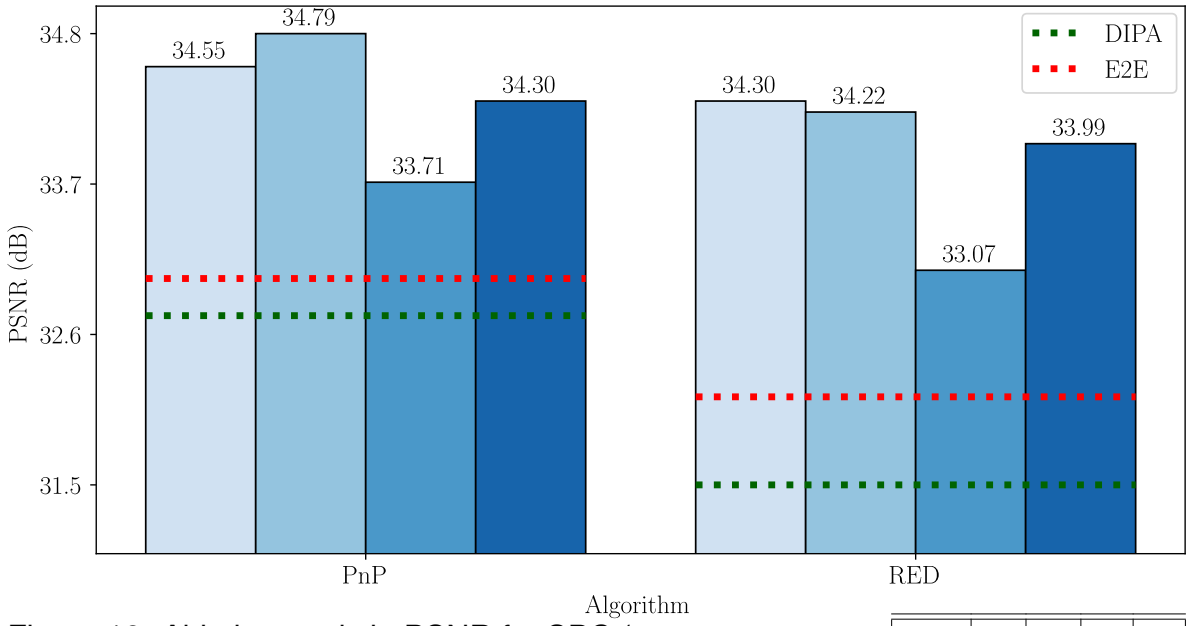
Learned	Method	Formulation	PSNR	Params	Ratio
✗	Baseline	$\mathbf{P} = \mathbf{I}_n$	22.36	0	0
✗	Hessian	$\mathbf{P} = (\mathbf{H}^\top \mathbf{H})^{-1}$	22.94	0	0
✗	Polynomial	$\mathbf{P} = p(\mathbf{H}^\top \mathbf{H})$	22.01	5	6×10^{-6}
✓	Scalar step	$\mathbf{P_k} = p_k \mathbf{I_n}$	23.36	20	2×10^{-5}
✓	Pointwise	$\mathbf{P_k} = \mathbf{p_k} \odot \mathbf{x}$	21.28	$20k$	0.024
✓	Convolutional	$\mathbf{P_k} = \mathbf{p_k} * \mathbf{x}$	21.30	500	6×10^{-4}
✓	Full-linear	$\mathbf{P_k}$	26.19	$21M$	24.42
✓	DIPA (Ours)	$\mathbf{P} = \mathbf{P}^\star$	32.74	$1M$	1.22
✓	Nonlinear E2E	$\mathcal{P} = \mathcal{P}^\star(\cdot)$	$\underline{33.79}$	$858k$	1
✓	D^{2}GP (Proposed)	$\mathcal{P} = \mathcal{P}^\star(\cdot)$	**34.55**	$858k$	1

Table 6. SOTA comparison for SPC ($\gamma_s = 0.2$, $\gamma_t = 0.7$) with FISTA-PnP.

In Table 7 of MRI, an improvement of more than 1 dB is observed, with 2000 and 7.2 times fewer parameters, than the Full-linear and DIPA methods.

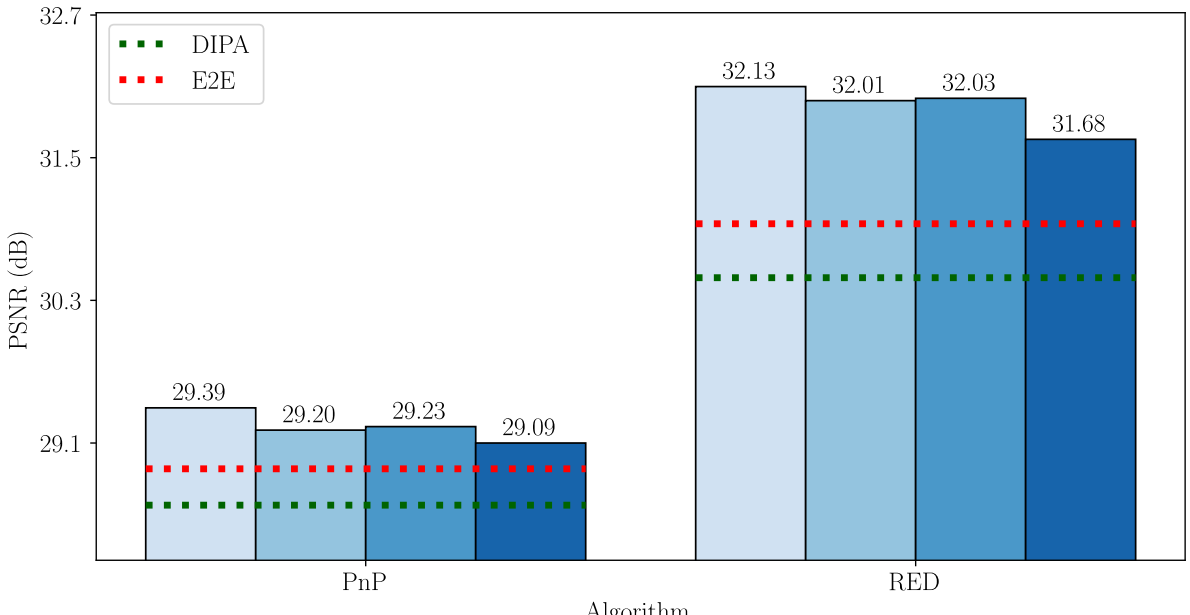
Learned	Method	Formulation	PSNR	Params	Ratio
✗	Baseline	$\mathbf{P} = \mathbf{I}_n$	25.77	0	0
✗	Hessian	$\mathbf{P} = (\mathbf{H}^\top \mathbf{H})^{-1}$	28.16	0	0
✗	Polynomial	$\mathbf{P} = p(\mathbf{H}^\top \mathbf{H})$	28.07	5	3×10^{-6}
✓	Scalar step	$\mathbf{P_k} = p_k \mathbf{I_n}$	27.88	20	1×10^{-5}
✓	Convolutional	$\mathbf{P_k} = \mathbf{p_k} * \mathbf{x}$	28.63	500	3×10^{-4}
✓	Pointwise	$\mathbf{P_k} = \mathbf{p_k} \odot \mathbf{x}$	27.59	$50k$	0.03
✓	Full-linear	$\mathbf{P_k}$	28.15	$125M$	72.8
✓	DIPA (Ours)	$\mathbf{P} = \mathbf{P}^\star$	28.56	$6M$	3.6
✓	Nonlinear E2E	$\mathcal{P} = \mathcal{P}^\star(\cdot)$	$\underline{28.87}$	$1.7M$	1
✓	D^{2}GP (Proposed)	$\mathcal{P} = \mathcal{P}^\star(\cdot)$	**29.39**	$1.7M$	1

Table 7. SOTA comparison for MRI ($AF_s = 5$, $AF_t = 1$) with FISTA-PnP.



$\mathcal{L}_{out}$	$\mathcal{L}_S$	$\mathcal{L}_S$	$\mathcal{L}_I$	$\mathcal{L}_I$
$\mathcal{R}_C$	✓	✗	✓	✗
Color				

Figure 18. Ablation study in PSNR for SPC ($\gamma_s = 0.2$, $\gamma_t = 0.7$) with PnP-FISTA and RED-FISTA for the D^{2}GP. Baseline SA is 23.24 dB and 22.93 dB. Author's own figure.



$\mathcal{L}_{out}$	$\mathcal{L}_S$	$\mathcal{L}_S$	$\mathcal{L}_I$	$\mathcal{L}_I$
$\mathcal{R}_C$	✓	✗	✓	✗
Color				

Figure 19. Ablation study in PSNR for MRI ($AF_s = 4$, $AF_t = 1$) with PnP-FISTA and RED-FISTA for the D^{2}GP. Baseline SA is 25.77 dB and 27.31 dB. Author's own figure.

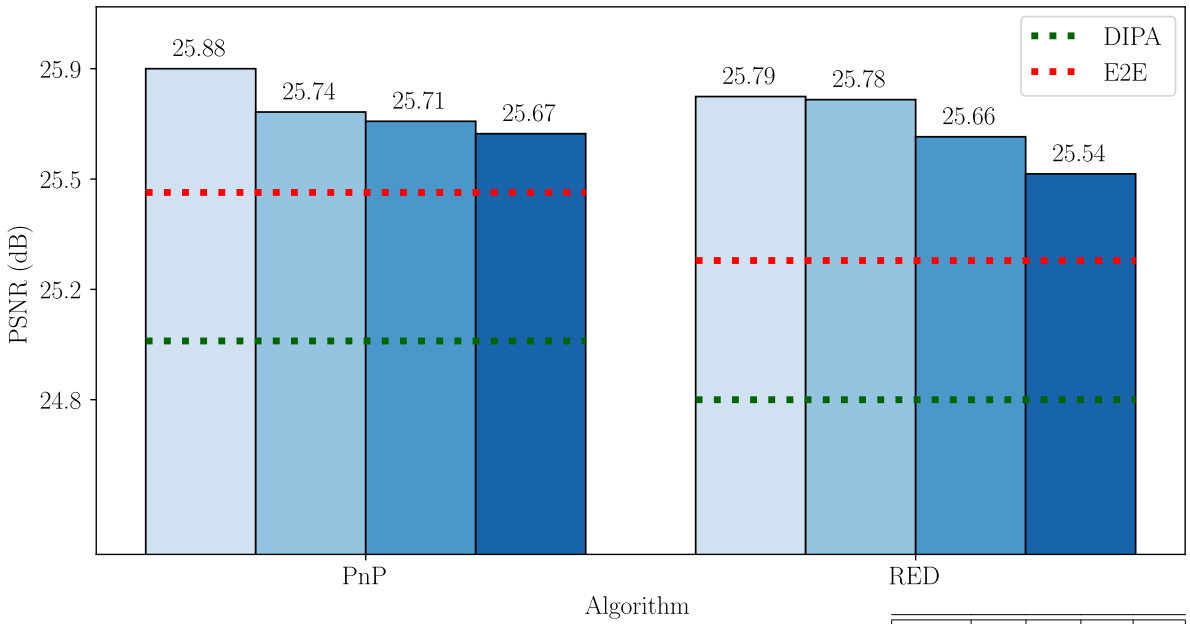


Figure 20. Ablation study in PSNR for SR ($RF_s = 4$, $RF_t = 1$) with PnP-FISTA and RED-FISTA for the D^{2}GP. Baseline SA is 11.10 dB and 10.88 dB. Author's own figure.

$\mathcal{L}_{out}$	$\mathcal{L}_S$	$\mathcal{L}_S$	$\mathcal{L}_I$	$\mathcal{L}_I$
$\mathcal{R}_C$	✓	✗	✓	✗
Color				

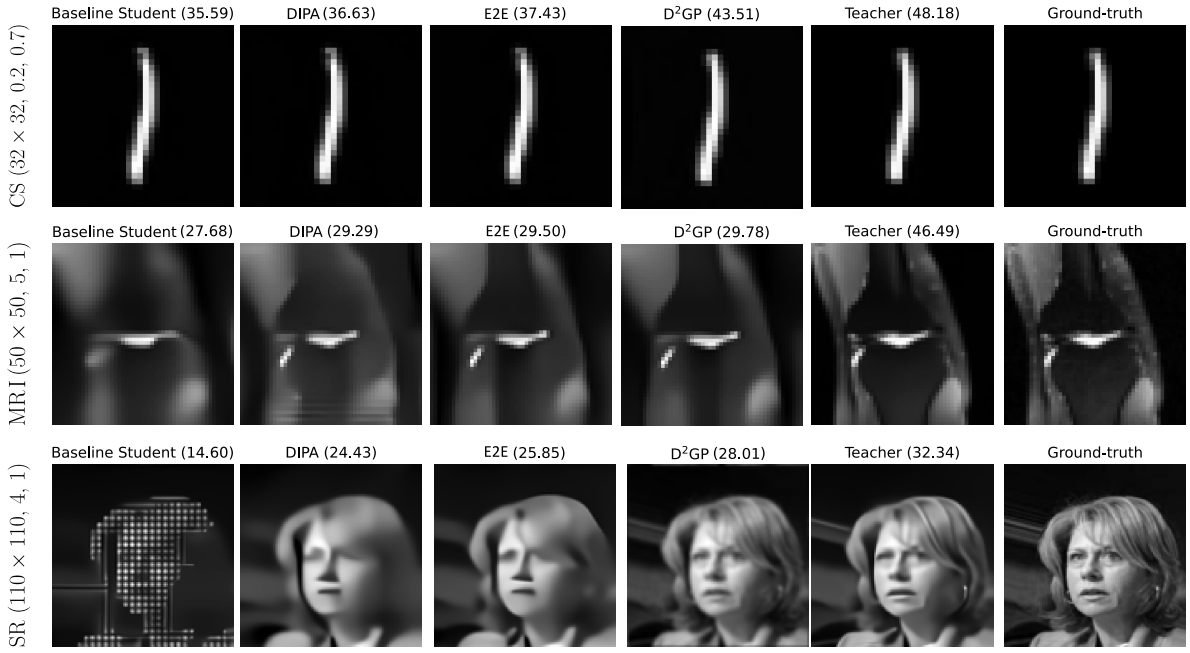


Figure 21. Visual results and PSNR for PnP-FISTA with D^{2}GP across different preconditioning methods. Author's own figure.

Finally, the largest reduction in trainable parameters is observed in SR, as it is gained by 3 dB and almost 1 dB, with 3410 and 170 times fewer parameters, than the Full-linear and DIPA methods, respectively. This last result evidences the invariance of D^{2}GP to the dimensionality of the images to be reconstructed, thanks to the internal use of convolutional networks in the ConvNeXt architecture. Additionally, although there are methods that require far fewer parameters than D^{2}GP, such as unlearnable or parameterized ones, their performance is very low, so although they are efficient, much is lost in terms of reconstruction quality.

Learned	Method	Formulation	PSNR	Params	Ratio
✗	Baseline	$\mathbf{P} = \mathbf{I}_n$	11.14	0	0
✗	Hessian	$\mathbf{P} = (\mathbf{H}^\top \mathbf{H})^{-1}$	11.75	0	0
✗	Polynomial	$\mathbf{P} = p(\mathbf{H}^\top \mathbf{H})$	11.81	5	6×10^{-6}
✓	Scalar step	$\mathbf{P_k} = p_k \mathbf{I_n}$	11.62	20	2×10^{-5}
✓	Pointwise	$\mathbf{P_k} = \mathbf{p_k} \odot \mathbf{x}$	11.19	$242k$	0.2818
✓	Convolutional	$\mathbf{P_k} = \mathbf{p_k} * \mathbf{x}$	17.55	500	6×10^{-4}
✓	Full-linear	$\mathbf{P_k}$	22.02	$2,928M$	3410
✓	DIPA (Ours)	$\mathbf{P} = \mathbf{P}^\star$	25.00	$146M$	170
✓	Nonlinear E2E	$\mathcal{P} = \mathcal{P}^\star(\cdot)$	$\underline{25.26}$	$858k$	1
✓	D^{2}GP (Proposed)	$\mathcal{P} = \mathcal{P}^\star(\cdot)$	**25.88**	$858k$	1

Table 8. SOTA comparison for SR ($RF_s = 4$, $RF_t = 1$) with FISTA-PnP.

3.5.5. Convergence and conditioning analysis In Fig. 22, we can observe the convergence of the reconstruction with respect to the GT, the convergence of the fidelity term, and the singular values of the Gram matrix for SPC and SR. In both cases, the proposed D^{2}GP method demonstrates the best convergence over the 20 iterations of the PnP-FISTA algorithm, requiring fewer iterations to obtain the performance of the other preconditioning methods. In terms of the fidelity term, it is observed that the proposed method is not the one with the best convergence, however, as already observed with the initial Measurement Augmentation technique, this is not bad. When comparing the convergence with the GT and the fidelity, it can be

seen that the other methods that had a lower fidelity, were stuck in a solution of a local minimum, which although it optimizes the fidelity, it is not the best reconstruction, this is due to the infinite possible solutions in the null space of these inverse problem. Finally, the singular values of the Gram matrix, which is a determinant term in the gradient of the algorithm, are analyzed. In this case, for the proposed method, due to its nonlinear nature, an approximation of its Jacobian is performed to calculate the condition number, as explained in Section 3.5.6. However, this approximation works in the space close to the reference point $\mathbf{x}_0$. Here it can be observed that there are methods that achieve a better condition number than the proposed method. However, this is not entirely related to the quality of the reconstruction, since it can be observed that although such methods as the Polynomial and the Hessian, have better stability, they do not manage to converge to the best solution. In conclusion, the proposed method achieves faster convergence, with better reconstruction quality, highlighting the low reliability of the fidelity term and the fact that a lower condition number does not always lead to a better reconstruction.

3.5.6. Local Linearization of the NPO via the Jacobian Let $\mathcal{P}_{\boldsymbol{\theta}^\star} : \mathbb{R}^n \to \mathbb{R}^n$ be the mapping defined by the pretrained nonlinear preconditioner. Assuming that $\mathcal{P}_{\boldsymbol{\theta}^\star}$ is differentiable in a neighborhood of a point $\mathbf{x}_0$, it can be linearized via its Jacobian $J_{\mathcal{P}_{\boldsymbol{\theta}^\star}}(\mathbf{x}_0)$, whose (i, j)-th entry is given by

$$\left(J_{\mathcal{P}_{\boldsymbol{\theta}^\star}}(\mathbf{x}_0)\right)_{ij} = \frac{\partial \left(\mathcal{P}_{\boldsymbol{\theta}^\star}(\mathbf{x}_0)\right)_i}{\partial x_j}.$$

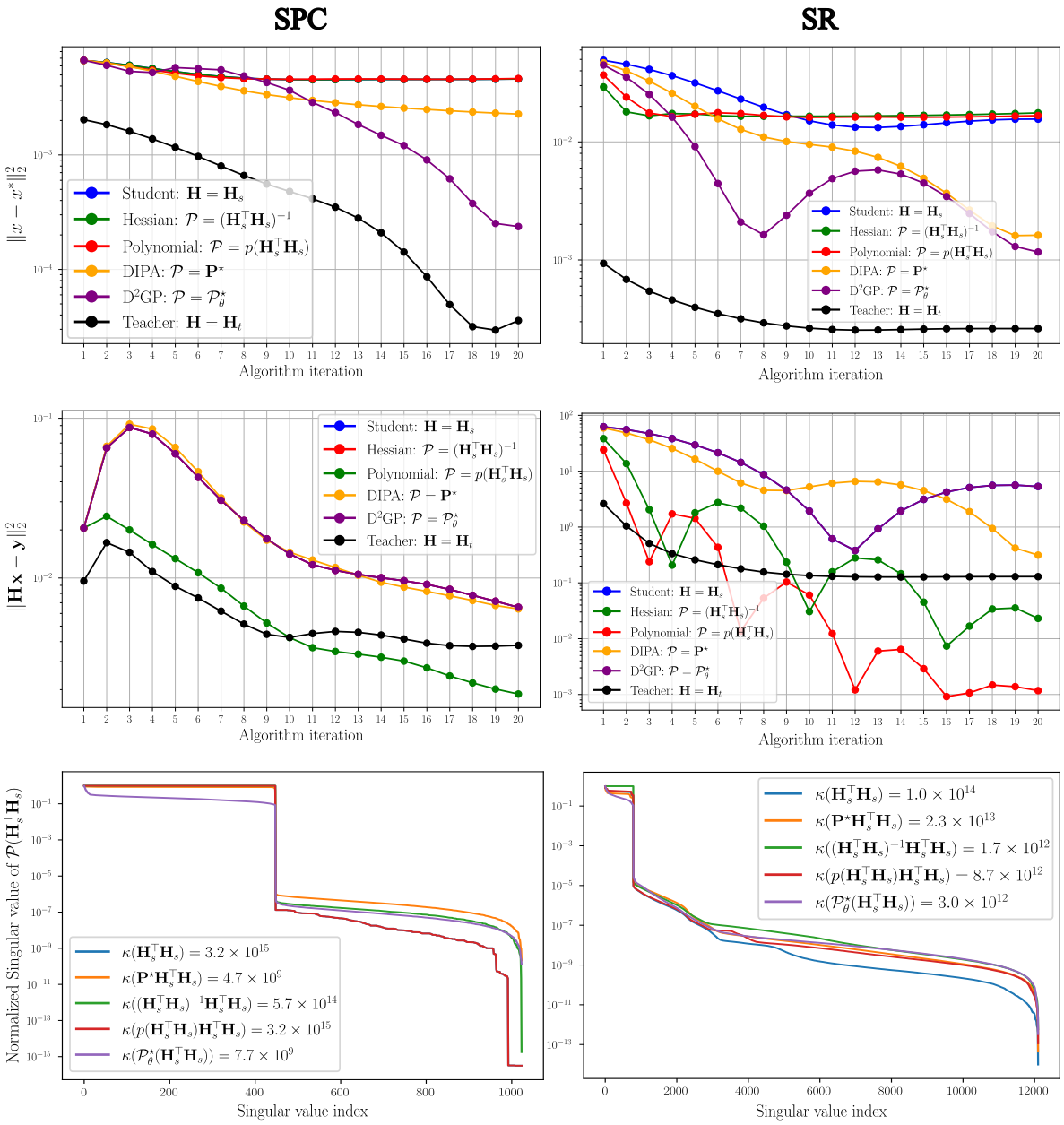


Figure 22. Reconstruction convergence, fidelity term convergence, and the Gram matrix's singular values for SPC and SR. Proposed and state-of-the-art preconditioning methods are validated. Author's own figure.

This Jacobian provides the best linear approximation of $\mathcal{P}_{\boldsymbol{\theta}^\star}$ around $\mathbf{x}_0$ [98,99]; that is, for sufficiently small perturbations $\boldsymbol{\delta} \in \mathbb{R}^n$,

$$\mathcal{P}_{\boldsymbol{\theta}^\star}(\mathbf{x}_0 + \boldsymbol{\delta}) \approx \mathcal{P}_{\boldsymbol{\theta}^\star}(\mathbf{x}_0) + J_{\mathcal{P}_{\boldsymbol{\theta}^\star}}(\mathbf{x}_0)\, \boldsymbol{\delta}.$$

This local linearization is the key to understanding how the nonlinear preconditioner behaves in the neighborhood of $\mathbf{x}_0$.

Finite Difference Approximation of the Jacobian

Because an analytic form of $J_{\mathcal{P}_{\boldsymbol{\theta}^\star}}(\mathbf{x}_0)$ is generally unavailable for deep neural networks, we approximate it using finite differences [100]. For a small $\epsilon > 0$, the partial derivative with respect to the j-th component is approximated by

$$\frac{\partial \left(\mathcal{P}_{\boldsymbol{\theta}^\star}(\mathbf{x}_0)\right)_i}{\partial x_j} \approx \frac{\left(\mathcal{P}_{\boldsymbol{\theta}^\star}(\mathbf{x}_0 + \epsilon\, \mathbf{e}_j)\right)_i - \left(\mathcal{P}_{\boldsymbol{\theta}^\star}(\mathbf{x}_0)\right)_i}{\epsilon},$$

where $\mathbf{e}_j$ is the j-th standard basis vector in $\mathbb{R}^n$. By performing this approximation for each component, we construct an approximate Jacobian matrix

$$J \approx \left[\frac{\mathcal{P}_{\boldsymbol{\theta}^\star}(\mathbf{x}_0 + \epsilon\, \mathbf{e}_j) - \mathcal{P}_{\boldsymbol{\theta}^\star}(\mathbf{x}_0)}{\epsilon}\right]_{j=1}^{n}.$$

This method is well-founded in numerical analysis [101]. With an optimal choice of ϵ,

---

[98] Arthur Jacot, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks". In: *Advances in neural information processing systems* 31 (2018).

[99] Maithra Raghu et al. "On the expressive power of deep neural networks". In: *international conference on machine learning*. PMLR. 2017, pp. 2847–2854.

[100] Bengt Fornberg. "Generation of finite difference formulas on arbitrarily spaced grids". In: *Mathematics of computation* 51.184 (1988), pp. 699–706.

[101] John E Dennis Jr and Robert B Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*. SIAM, 1996.

the approximation becomes as accurate as possible given the hardware precision.

Comparison with the Linear Case

In the linear preconditioning case, one analyzes the eigen- or singular value spectrum of $\mathbf{P}^\star\mathbf{H}^\top\mathbf{H}$ to assess convergence and stability in the preconditioned FISTA algorithm. Here, by approximating the Jacobian $J_{\mathcal{P}_{\boldsymbol{\theta}^\star}}(\mathbf{x}_0)$ of the nonlinear preconditioner, we obtain a local linear model of $\mathcal{P}_{\boldsymbol{\theta}^\star}$ at $\mathbf{x}_0$. This local model allows us to compute singular values that can be directly compared with the spectrum obtained from the linear operator $\mathbf{P}^\star$.

Given that $\mathcal{P}_{\boldsymbol{\theta}^\star}(\cdot)$ does not possess a fixed matrix representation due to its inherent nonlinearity, this approach of approximating the Jacobian is the only feasible method to extract local spectral information, which in turn allows for a direct comparison with the linear preconditioner $\mathbf{P}^\star$.

By approximating the Jacobian $J_{\mathcal{P}_{\boldsymbol{\theta}^\star}}(\mathbf{x}_0)$ via finite differences and performing an SVD, we can extract the singular values that characterize the local behavior of the pretrained nonlinear preconditioner $\mathcal{P}_{\boldsymbol{\theta}^\star}$ [102]. These singular values serve as indicators of local amplification and stability, with the condition number offering a quantitative measure of sensitivity. This analysis provides a rigorous, mathematically sound framework for comparing the nonlinear preconditioner with the traditional linear preconditioner $\mathbf{P}^\star$, as used in the preconditioned FISTA algorithm [103].

3.5.7. Visual representation of the NPO In Fig. 23, the linear representation of the learned preconditioning operator $\mathcal{P}_{\boldsymbol{\theta}^\star}$ can be observed. The patterns demonstrate the adaptation of the proposed method to each task. Dominance of the values on the diagonal is also observed, similar to what was obtained with the DIPA method.

---

[102] Andreas Griewank and Andrea Walther. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. SIAM, 2008.

[103] Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

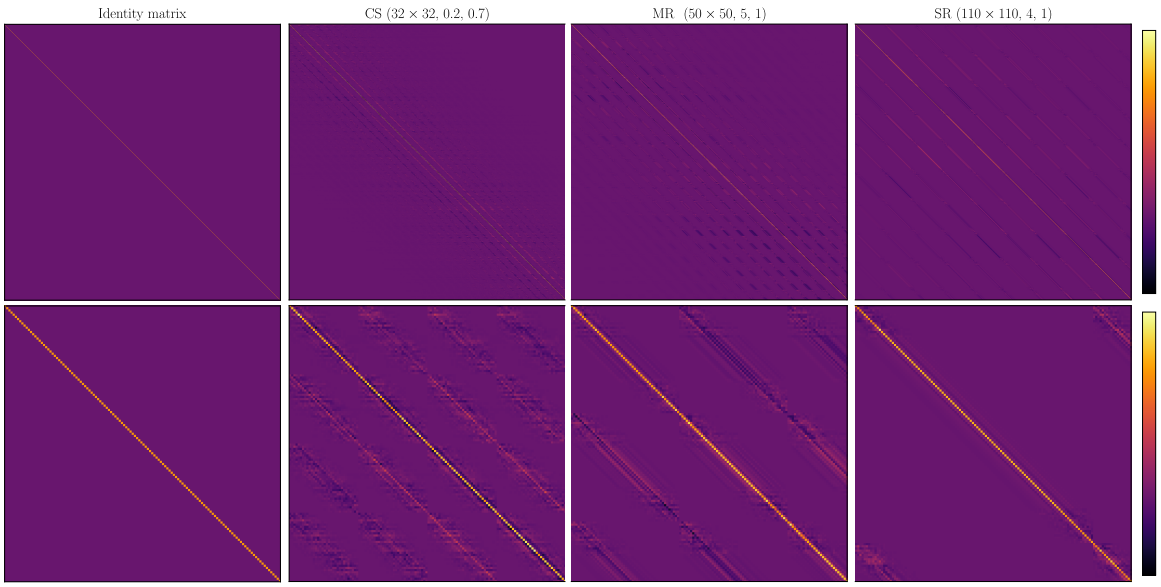


Figure 23. Linear representation of the learned preconditioning operator $\mathcal{P}_{\boldsymbol{\theta}^\star}$ using approximation from Sec. 3.5.6. **First row:** 1024×1024 zoomed version. **Second row:** 128×128 zoomed version. **Title format:** Imaging task (Spatial resolution, Student information, Teacher information). Author's own figure.

4. Discussion and Future Work

This thesis presents two complementary approaches for gradient preconditioning in imaging inverse problems. On one hand, the **Gradient Preconditioning via Measurement Augmentation** strategy, which can be interpreted as a form of preconditioning. In this approach, a neural network generates additional measurements that are concatenated with the original ones. The Jacobian of the augmented measurements then modifies the initial gradient, which results in improved convergence in fewer iterations. On the other hand, **Gradient Preconditioning via Knowledge Distillation**, can be linear or nonlinear. The *Linearly Distilled Gradient Preconditioned Algorithms* (DIPA) offers an interpretable preconditioning operator that allows for a clear analysis of its behavior. Although its performance is not as high as that of the non-linear method, there is significant potential to improve DIPA further by integrating low-rank optimization techniques (e.g., LORA). Such enhancements could reduce the number of parameters while maintaining its inherent interpretability, addressing one of the main limitations of neural network–based approaches. The proposed *Deep Distillation for Non-Linear Gradient Preconditioning* (D^2GP) method leverages a non-linear, black-box approach using convolutional architectures. D^2GP outperforms end-to-end methods for designing the preconditioning operator by integrating a teacher algorithm. Although the teacher does not have access to the ground truth, their guidance allows the student to achieve superior reconstructions.

Future research should explore alternative loss functions and knowledge distillation strategies, such as self-distillation or online distillation, to further enhance these methods. Moreover, extending the validation of these techniques to higher-resolution imaging tasks and implementing them in laboratory settings will be essential. Finally, the design of the teacher algorithm remains an intriguing open question that could further refine the quality of the preconditioning operator.

5. Conclusions

This research work presents a comprehensive framework for tackling imaging inverse problems by integrating gradient preconditioning. The initially proposed measurement augmentation technique has proven effective by generating complementary measurements that refine the initial gradient and accelerate convergence. The linear and interpretable DIPA approach provides valuable insights into the preconditioning process. With further enhancements through low-rank optimization (LORA), DIPA has the potential to remain competitive while retaining its interpretability. The proposed D^2GP method demonstrates that incorporating a teacher algorithm to guide the student not only enhances reconstruction quality but also surpasses conventional end-to-end approaches for designing the preconditioning operator. Its robustness to image dimensionality—achieved through the use of convolutional architectures—confirms the method's practical applicability across various imaging modalities.

Overall, the contributions of this work advance the state-of-the-art in gradient preconditioning and offer several promising avenues for future exploration. These include the investigation of novel loss functions, alternative distillation techniques, and the continued refinement of the teacher algorithm design—all of which are key to developing more efficient, robust, and interpretable solutions for imaging inverse problems.

BIBLIOGRAPHY

Aggarwal, Hemant K, Merry P Mani, and Mathews Jacob. “MoDL: Model-based deep learning architecture for inverse problems”. In: *IEEE transactions on medical imaging* 38.2 (2018), pp. 394–405 (cit. on p. 27).

Arbelaez, Pablo et al. “Contour detection and hierarchical image segmentation”. In: *IEEE transactions on pattern analysis and machine intelligence* 33.5 (2010), pp. 898–916 (cit. on p. 59).

Baldi, Pierre. “Autoencoders, unsupervised learning, and deep architectures”. In: *Proceedings of ICML workshop on unsupervised and transfer learning*. JMLR Workshop and Conference Proceedings. 2012, pp. 37–49 (cit. on p. 29).

Beck, Amir and Marc Teboulle. “A fast iterative shrinkage-thresholding algorithm for linear inverse problems”. In: *SIAM journal on imaging sciences* 2.1 (2009), pp. 183–202 (cit. on pp. 22, 24, 25, 54).

Benning, Martin and Martin Burger. “Modern regularization methods for inverse problems”. In: *Acta numerica* 27 (2018), pp. 1–111 (cit. on p. 26).

Benzi, Michele. “Preconditioning techniques for large linear systems: a survey”. In: *Journal of computational Physics* 182.2 (2002), pp. 418–477 (cit. on p. 17).

Bertero, Mario, Patrizia Boccacci, and Christine De Mol. *Introduction to inverse problems in imaging*. CRC press, 2021 (cit. on p. 21).

Boyd, Stephen et al. “Distributed optimization and statistical learning via the alternating direction method of multipliers”. In: *Foundations and Trends® in Machine learning* 3.1 (2011), pp. 1–122 (cit. on pp. 23, 25, 42).

Burger, Harold C, Christian J Schuler, and Stefan Harmeling. “Image denoising: Can plain neural networks compete with BM3D?” In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 2392–2399 (cit. on p. 26).

Caraba, Elena. “Preconditioned conjugate gradient algorithm”. In: (2008) (cit. on p. 32).

Chan, Stanley H, Xiran Wang, and Omar A Elgendy. “Plug-and-play ADMM for image restoration: Fixed-point convergence and applications”. In: *IEEE Transactions on Computational Imaging* 3.1 (2016), pp. 84–98 (cit. on pp. 42, 52).

Cline, Alan K et al. “An estimate for the condition number of a matrix”. In: *SIAM Journal on Numerical Analysis* 16.2 (1979), pp. 368–375 (cit. on p. 30).

Dabov, Kostadin et al. “Image denoising with block-matching and 3D filtering”. In: *Image processing: algorithms and systems, neural networks, and machine learning*. Vol. 6064. SPIE. 2006, pp. 354–365 (cit. on p. 26).

Dassios, Ioannis, Kimon Fountoulakis, and Jacek Gondzio. “A preconditioner for a primal-dual newton conjugate gradient method for compressed sensing problems”. In: *SIAM Journal on Scientific Computing* 37.6 (2015), A2783–A2812 (cit. on pp. 34, 68).

Delbracio, Mauricio and Peyman Milanfar. “Inversion by direct iteration: An alternative to denoising diffusion for image restoration”. In: *arXiv preprint arXiv:2303.11435* (2023) (cit. on pp. 66, 67).

Deng, Li. “The mnist database of handwritten digit images for machine learning research [best of the web]”. In: *IEEE signal processing magazine* 29.6 (2012), pp. 141–142 (cit. on pp. 43, 55).

Dennis Jr, John E and Robert B Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*. SIAM, 1996 (cit. on p. 75).

Dosovitskiy, Alexey et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020) (cit. on p. 66).

Duarte, Marco F et al. “Single-pixel imaging via compressive sampling”. In: *IEEE signal processing magazine* 25.2 (2008), pp. 83–91 (cit. on pp. 20, 49).

Ehrhardt, Matthias J, Patrick Fahy, and Mohammad Golbabaee. “Learning preconditioners for inverse problems”. In: *arXiv e-prints* (2024), arXiv–2406 (cit. on pp. 59, 68).

Erdogan, Atakan. *ConvNeXt: Next Generation of Convolutional Networks*. `https://medium.com/@atakanerdogan305/convnext-next-generation-of-convolutional-networks-325607a08c46`. [Accessed: 17-Mar-2025]. 2021 (cit. on p. 67).

Fessler, Jeffrey A and Scott D Booth. “Conjugate-gradient preconditioning methods for shift-variant PET image reconstruction”. In: *IEEE transactions on image processing* 8.5 (1999), pp. 688–699 (cit. on pp. 34, 68).

Fornberg, Bengt. “Generation of finite difference formulas on arbitrarily spaced grids”. In: *Mathematics of computation* 51.184 (1988), pp. 699–706 (cit. on p. 75).

Gander, Martin J. “On the origins of linear and non-linear preconditioning”. In: *Domain decomposition methods in science and engineering XXIII*. Springer. 2017, pp. 153–161 (cit. on pp. 17, 35, 43).

Garber, Tomer and Tom Tirer. “Image Restoration by Denoising Diffusion Models with Iteratively Preconditioned Guidance”. In: *arXiv preprint arXiv:2312.16519* (2023) (cit. on p. 33).

— “Image restoration by denoising diffusion models with iteratively preconditioned guidance”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 25245–25254 (cit. on p. 59).

Gibson, Graham M, Steven D Johnson, and Miles J Padgett. “Single-pixel imaging 12 years on: a review”. In: *Optics express* 28.19 (2020), pp. 28190–28208 (cit. on p. 39).

Godfrey, Andrew. “Steps toward a robust preconditioning”. In: *32nd Aerospace Sciences Meeting and Exhibit*. 1994, p. 520 (cit. on p. 17).

Golub, Gene H, Per Christian Hansen, and Dianne P O’Leary. “Tikhonov regularization and total least squares”. In: *SIAM journal on matrix analysis and applications* 21.1 (1999), pp. 185–194 (cit. on p. 25).

Golub, Gene H and Charles F Van Loan. *Matrix computations*. JHU press, 2013 (cit. on p. 76).

Goodfellow, Ian et al. “Generative adversarial networks”. In: *Communications of the ACM* 63.11 (2020), pp. 139–144 (cit. on p. 29).

Gou, Jianping et al. “Knowledge distillation: A survey”. In: *International Journal of Computer Vision* 129.6 (2021), pp. 1789–1819 (cit. on p. 36).

Griewank, Andreas and Andrea Walther. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. SIAM, 2008 (cit. on p. 76).

Gualdrón-Hurtado, Romario, Henry Arguello, and Jorge Bacca. “Deep Learned Non-Linear Propagation Model Regularizer for Compressive Spectral Imaging”. In: *IEEE Transactions on Computational Imaging* 10 (2024), pp. 1016–1025. DOI: 10.1109/TCI.2024.3422900 (cit. on pp. 13, 28).

Gualdrón-Hurtado, Romario et al. “Improving Compressive Imaging Recovery via Measurement Augmentation”. In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2025, pp. 1–5. DOI: 10.1109/ICASSP49660.2025.10888734 (cit. on pp. 13, 39).

Gualdrón-Hurtado, Romario et al. “Learning Point Spread Function Invertibility Assessment for Image Deconvolution”. In: *2024 32nd European Signal Processing Conference (EUSIPCO)*. 2024, pp. 501–505. DOI: 10.23919/EUSIPCO63174.2024.10715342 (cit. on p. 14).

Gurney, Kevin. *An introduction to neural networks*. CRC press, 2018 (cit. on p. 28).

Haber, Eldad, Uri M Ascher, and Doug Oldenburg. “On optimization techniques for solving nonlinear inverse problems”. In: *Inverse problems* 16.5 (2000), p. 1263 (cit. on p. 34).

Hartmann, Morghan, Hasan Farooq, and Ali Imran. “Distilled deep learning based classification of abnormal heartbeat using ECG data through a low cost edge

device". In: *2019 IEEE symposium on computers and communications (ISCC)*. IEEE. 2019, pp. 1068–1071 (cit. on p. 36).

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. *Distilling the Knowledge in a Neural Network*. 2015. arXiv: 1503.02531 [stat.ML] (cit. on p. 48).

Hu, W et al. "Preconditioned non-linear conjugate gradient method for frequency domain full-waveform seismic inversion". In: *Geophysical Prospecting* 59.3 (2011), pp. 477–491 (cit. on p. 33).

Hwang, Feng-Nan and Xiao-Chuan Cai. "Improving robustness and parallel scalability of Newton method through nonlinear preconditioning". In: *Domain decomposition methods in science and engineering*. Springer, 2005, pp. 201–208 (cit. on p. 35).

Iyer, Siddharth S et al. "Polynomial preconditioners for regularized linear inverse problems". In: *SIAM Journal on Imaging Sciences* 17.1 (2024), pp. 116–146 (cit. on pp. 36, 68).

Jacome, Roman et al. "Learning to Reconstruct Signals With Inexact Sensing Operator via Knowledge Distillation". In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2025, pp. 1–5. DOI: 10.1109/ICASSP49660.2025.10887652 (cit. on pp. 14, 38).

Jacot, Arthur, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks". In: *Advances in neural information processing systems* 31 (2018) (cit. on p. 75).

Jerez, Andrés, Miguel Márquez, and Henry Arguello. “Adaptive coded aperture design for compressive computed tomography”. In: *Journal of Computational and Applied Mathematics* 384 (2021), p. 113174 (cit. on p. 16).

Jin, Bangti, Peter Maaß, and Otmar Scherzer. “Sparsity regularization in inverse problems”. In: *Inverse Problems* 33.6 (2017) (cit. on p. 26).

Johnson, Olin G, Charles A Micchelli, and George Paul. “Polynomial preconditioners for conjugate gradient calculations”. In: *SIAM Journal on Numerical Analysis* 20.2 (1983), pp. 362–376 (cit. on pp. 36, 68).

Jones, Andrew Charles. *Conjugate Gradients*. `https://andrewcharlesjones.github.io/journal/conjugate-gradients.html`. [Accessed: 17-Mar-2025]. 2023 (cit. on p. 30).

Kanwal, Ayesha et al. “A hybrid framework for detection of autism using ConvNeXt-T and embedding clusters”. In: *The Journal of Supercomputing* 80.6 (2024), pp. 8156–8178 (cit. on p. 67).

Karl, W Clem et al. “The Foundations of Computational Imaging: A signal processing perspective”. In: *IEEE Signal Processing Magazine* 40.5 (2023), pp. 40–53 (cit. on p. 20).

Kershaw, David S. “The incomplete Cholesky—conjugate gradient method for the iterative solution of systems of linear equations”. In: *Journal of computational physics* 26.1 (1978), pp. 43–65 (cit. on p. 33).

Kettler, Rob. “Analysis and comparison of relaxation schemes in robust multigrid and preconditioned conjugate gradient methods”. In: *Multigrid Methods: Proceedings*

of the Conference Held at Köln-Porz, November 23–27, 1981. Springer. 2006, pp. 502–534 (cit. on p. 17).

Kingma, Diederik P. and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015 (cit. on pp. 44, 52, 55).

Knoll, Florian et al. “FastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning”. en. In: *Radiol. Artif. Intell.* 2.1 (Jan. 2020), e190007 (cit. on p. 54).

LeCun, Yann et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324 (cit. on p. 66).

Lee, Youngdo. *[DL] Natural Gradient*. `https://leeyngdo.github.io/blog/deep-learning/2024-02-01-natural-gradient`. [Accessed: 17-Mar-2025]. 2024 (cit. on p. 30).

Li, Xi-Lin. “Preconditioned stochastic gradient descent”. In: *IEEE transactions on neural networks and learning systems* 29.5 (2017), pp. 1454–1466 (cit. on pp. 32, 35).

Liu, Lulu, David E Keyes, and Rolf Krause. “A note on adaptive nonlinear preconditioning techniques”. In: *SIAM Journal on Scientific Computing* 40.2 (2018), A1171–A1186 (cit. on p. 34).

Liu, Po-Yu and Edmund Y Lam. “Image reconstruction using deep learning”. In: *arXiv preprint arXiv:1809.10410* (2018) (cit. on p. 29).

Liu, Zhuang et al. “A convnet for the 2020s”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 11976–11986 (cit. on p. 66).

Liu, Ziwei et al. “Deep Learning Face Attributes in the Wild”. In: *Proceedings of International Conference on Computer Vision (ICCV)*. 2015 (cit. on pp. 55, 57).

Loshchilov, I. “Decoupled weight decay regularization”. In: *arXiv preprint arXiv:1711.05101* (2017) (cit. on pp. 52, 54).

Martinez, Emmanuel et al. “Compressive Imaging Reconstruction via Conditional Diffusion Model With Augmented Measurements”. In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2025, pp. 1–5. DOI: 10.1109/ICASSP49660.2025.10889114 (cit. on p. 14).

Narayanan, Arvind and Vitaly Shmatikov. “Robust de-anonymization of large sparse datasets”. In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE. 2008, pp. 111–125 (cit. on pp. 17, 35).

Nielsen, Bjørn Fredrik and Kent-Andre Mardal. “Efficient preconditioners for optimality systems arising in connection with inverse problems”. In: *SIAM Journal on Control and Optimization* 48.8 (2010), pp. 5143–5177 (cit. on p. 30).

Ongie, Gregory et al. “Deep learning techniques for inverse problems in imaging”. In: *IEEE Journal on Selected Areas in Information Theory* 1.1 (2020), pp. 39–56 (cit. on p. 27).

Paszke, Adam et al. “Pytorch: An imperative style, high-performance deep learning library”. In: *Advances in neural information processing systems* 32 (2019) (cit. on p. 42).

Pruessmann, Klaas P et al. “Coil sensitivity encoding for fast MRI”. In: *Proceedings of the ISMRM 6th Annual Meeting, Sydney*. Vol. 1998. 1998 (cit. on p. 49).

Qin, Dian et al. “Efficient medical image segmentation based on knowledge distillation”. In: *IEEE Transactions on Medical Imaging* 40.12 (2021), pp. 3820–3831 (cit. on p. 36).

Raghu, Maithra et al. “On the expressive power of deep neural networks”. In: *international conference on machine learning*. PMLR. 2017, pp. 2847–2854 (cit. on p. 75).

Rajasegaran, Jathushan et al. “Self-supervised knowledge distillation for few-shot learning”. In: *arXiv preprint arXiv:2006.09785* (2020) (cit. on p. 37).

Romano, Yaniv, Michael Elad, and Peyman Milanfar. “The little engine that could: Regularization by denoising (RED)”. In: *SIAM Journal on Imaging Sciences* 10.4 (2017), pp. 1804–1844 (cit. on p. 27).

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241 (cit. on p. 66).

Rosenblatt, Frank. “The perceptron: a probabilistic model for information storage and organization in the brain.” In: *Psychological review* 65.6 (1958), p. 386 (cit. on p. 65).

Rutishauser, Heinz. “The Jacobi method for real symmetric matrices”. In: *Numerische Mathematik* 9.1 (1966), pp. 1–10 (cit. on p. 33).

Schmidt, Mark. “Least squares optimization with L1-norm regularization”. In: *CS542B Project Report* 504.2005 (2005), pp. 195–221 (cit. on p. 24).

Shanno, David F and Kang Hoh Phua. “Matrix conditioning and nonlinear optimization”. In: *Mathematical Programming* 14 (1978), pp. 149–160 (cit. on pp. 16, 20).

Stanton, Samuel et al. “Does knowledge distillation really work?” In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 6906–6919 (cit. on p. 37).

Stewart, Gilbert W. “On the early history of the singular value decomposition”. In: *SIAM review* 35.4 (1993), pp. 551–566 (cit. on p. 31).

Strong, David and Tony Chan. “Edge-preserving and scale-dependent properties of total variation regularization”. In: *Inverse problems* 19.6 (2003), S165 (cit. on p. 25).

Suarez-Rodriguez, Leon, Roman Jacome, and Henry Arguello. “Distilling Knowledge for Designing Computational Imaging Systems”. In: *arXiv preprint arXiv:2501.17898* (2025) (cit. on pp. 38, 48).

— “Highly Constrained Coded Aperture Imaging Systems Design Via a Knowledge Distillation Approach”. In: *arXiv e-prints* (2024), arXiv–2406 (cit. on pp. 37, 38).

Tachella, Julián, Dongdong Chen, and Mike Davies. “Sensing theorems for unsupervised learning in linear inverse problems”. In: *Journal of Machine Learning Research* 24.39 (2023), pp. 1–45 (cit. on p. 29).

— “Unsupervised learning from incomplete measurements for inverse problems”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 4983–4995 (cit. on p. 29).

Tachella, Julian et al. *DeepInverse: A deep learning framework for inverse problems in imaging*. Version latest. June 2023. DOI: 10.5281/zenodo.7982256 (cit. on p. 43).

— *DeepInverse: A deep learning framework for inverse problems in imaging*. Version latest. June 2023. DOI: 10.5281/zenodo.7982256 (cit. on p. 54).

Tan, Hong Ye et al. "Provably convergent plug-and-play quasi-Newton methods". In: *SIAM Journal on Imaging Sciences* 17.2 (2024), pp. 785–819 (cit. on pp. 36, 61, 68).

Tian, Chunwei et al. "Deep learning on image denoising: An overview". In: *Neural Networks* 131 (2020), pp. 251–275 (cit. on p. 29).

Tselepidis, Nikolaos, Jonas Kohler, and Antonio Orvieto. "Two-level K-FAC preconditioning for deep learning". In: *arXiv preprint arXiv:2011.00573* (2020) (cit. on p. 33).

Uhlmann, Gunther. "Inverse problems: seeing the unseen". In: *Bulletin of Mathematical Sciences* 4 (2014), pp. 209–279 (cit. on p. 28).

Urrea, Sergio et al. "Optical Solutions for Spectral Imaging Inverse Problems with a Shift-Variant System." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 4157–4164 (cit. on p. 16).

Venkatakrishnan, Singanallur V., Charles A. Bouman, and Brendt Wohlberg. "Plug-and-Play priors for model based reconstruction". In: *2013 IEEE Global Conference on Signal and Information Processing*. 2013, pp. 945–948. DOI: 10.1109/GlobalSIP.2013.6737048 (cit. on p. 26).

Vlaardingerbroek, Marinus T and Jacques A Boer. *Magnetic resonance imaging: theory and practice*. Springer Science & Business Media, 2013 (cit. on p. 20).

Vogel, Curtis R. *Computational methods for inverse problems*. SIAM, 2002 (cit. on p. 16).

Wang, Zhihao, Jian Chen, and Steven CH Hoi. “Deep learning for image super-resolution: A survey”. In: *IEEE transactions on pattern analysis and machine intelligence* 43.10 (2020), pp. 3365–3387 (cit. on p. 29).

Woo, Sanghyun et al. “Cbam: Convolutional block attention module”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 3–19 (cit. on p. 66).

Yang, Jianchao et al. “Image super-resolution via sparse representation”. In: *IEEE transactions on image processing* 19.11 (2010), pp. 2861–2873 (cit. on pp. 20, 50).

Yu, Wen-Kai. “Super sub-Nyquist single-pixel imaging by means of cake-cutting Hadamard basis sort”. In: *Sensors* 19.19 (2019), p. 4122 (cit. on p. 43).

Yuan, Qiangqiang et al. “A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11.3 (2018), pp. 978–989 (cit. on p. 66).

Zhang, Kai et al. “Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising”. In: *IEEE transactions on image processing* 26.7 (2017), pp. 3142–3155 (cit. on p. 43).

Zhu, Bo et al. “Image reconstruction by domain-transform manifold learning”. In: *Nature* 555.7697 (2018), pp. 487–492 (cit. on p. 29).