DATA-DRIVEN PHASE RETRIEVAL USING DEEP GENERATIVE MODELS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

MEHMET ONURCAN KAYA

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

JUNE 2024

Approval of the thesis:

DATA-DRIVEN PHASE RETRIEVAL USING DEEP GENERATIVE MODELS

submitted by **MEHMET ONURCAN KAYA** in partial fulfillment of the requirements for the degree of **Master of Science in Electrical and Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Naci Emre Altun
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. İlkay Ulusoy
Head of Department, **Electrical and Electronics Engineering** _____

Assoc. Prof. Dr. S. Figen Öktem
Supervisor, **Electrical and Electronics Engineering, METU** _____

Examining Committee Members:

Assoc. Prof. Dr. Elif Vural
Electrical and Electronics Engineering, METU _____

Assoc. Prof. Dr. S. Figen Öktem
Electrical and Electronics Engineering, METU _____

Prof. Dr. Tolga Çukur
Electrical and Electronics Engineering, Bilkent University _____

Date: 25.06.2024

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Mehmet Onurcan Kaya

Signature :

ABSTRACT

DATA-DRIVEN PHASE RETRIEVAL USING DEEP GENERATIVE MODELS

Kaya, Mehmet Onurcan
M.S., Department of Electrical and Electronics Engineering
Supervisor: Assoc. Prof. Dr. S. Figen Öktem

June 2024, 107 pages

This thesis addresses the nonlinear inverse problem of phase retrieval, which is the process of recovering a signal from the magnitude of its Fourier transform, a fundamental challenge in fields such as electron microscopy, crystallography, astronomy, and optical imaging. Classical phase retrieval techniques face limitations in robustness, noise sensitivity, and computational efficiency. To overcome these limitations, this work develops novel data-driven phase retrieval methods by exploiting advanced deep generative models. Firstly, we present a phase retrieval approach leveraging Langevin dynamics within diffusion models. This approach utilizes two different deep learning pipelines, namely prNet-Small and prNet-Large, and carefully balances the perceptual quality-distortion tradeoff. While we favor minimal distortion, we also aim to create high-perceptual quality images. Secondly, we use the Inversion by Direct Denoising (InDI) framework to solve the Fourier phase retrieval problem. The developed method also employs advanced initialization strategies and ensembling techniques, resulting in improved training efficiency and better image quality compared to traditional methods. Thirdly, we extend the Denoising Diffusion Restoration Models (DDRM) for phase retrieval by combining with the Hybrid Input-Output (HIO)

method. This approach utilizes pretrained unconditional diffusion models. Overall, this thesis demonstrates that exploiting the score/diffusion-based framework significantly improves the solution of the phase retrieval problem by enabling unprecedented image quality, better noise robustness, and higher computational speed and efficiency. These advancements have a broad impact on computational imaging and various scientific and engineering applications.

Keywords: Phase Retrieval, Diffusion, Generative Models, Deep Learning, Image Reconstruction

ÖZ

DERİN ÜRETİCİ MODELLER İLE VERİ GÜDÜMLÜ FAZ GERİ KAZANIMI

Kaya, Mehmet Onurcan
Yüksek Lisans, Elektrik ve Elektronik Mühendisliği Bölümü
Tez Yöneticisi: Doç. Dr. S. Figen Öktem

Haziran 2024 , 107 sayfa

Bu tez, elektron mikroskopisi, kristalografi, astronomi ve optik görüntüleme gibi alanlarda önemli uygulamaları olan faz geri kazanımı isimli doğrusal olmayan bir ters problemi ele almaktadır. Bir sinyalin Fourier dönüşümünün sadece genlik değerlerinden yola çıkarak bu sinyalin algoritmik olarak geri kazanılması amaçlanan bu problemde klasik olarak kullanılan yöntemler sağlamlık, gürültü hassasiyeti ve hesaplama verimliliği açısından oldukça sınırlıdırlar. Bu sınırlamaların üstesinden gelmek için bu çalışmada ileri düzey derin üretici modeller yardımıyla veri güdümlü yeni faz geri kazanım teknikleri geliştirilmiştir. İlk olarak, Langevin dinamiği tekniğini baz alan bir faz geri kazanım yaklaşımı anlatılmaktadır. Bu yöntem, prNet-Small ve prNet-Large ismindeki derin öğrenme yapılarını kullanmakta ve algısal kalite-bozulma dengesini göz önünde bulundurarak minimum bozulma yanında yüksek algısal kalitede geri-çatım elde etmektedir. İkinci olarak, Doğrudan Gürültü Giderme ile Ters Çevirme (InDI) çerçevesi, Fourier faz geri kazanımı için denenmektedir. Bu yöntem, ayrıca gelişmiş başlangıç stratejileri ve birleştirme teknikleri kullanarak, geleneksel yöntemlere kıyasla eğitim verimliliğini ve görüntü kalitesini artırmaktadır. Son olarak, Gü-

rültü Giderme Difüzyon Geri Kazanım Modelleri (DDRM) faz geri kazanımı problemine Hibrit Giriş-Çıkış (HIO) yöntemi kullanılarak genişletilmektedir. Bu yaklaşım, faz geri kazanım performansını artırmak için önceden eğitilmiş koşulsuz difüzyon modellerini kullanmaktadır. Genel olarak, bu tez, skor/difüzyon tabanlı çerçeveyi entegre etmenin faz geri kazanımının görsel performansını önemli ölçüde iyileştirdiğini, daha iyi gürültü dayanıklılığı ve daha iyi hesaplama başarımı sunduğunu göstermektedir. Bu gelişmeler, hesaplamalı görüntüleme ve çeşitli bilim ve mühendislik uygulamaları için geniş kapsamlı etkilere sahiptir.

Anahtar Kelimeler: Faz Geri Kazanımı, Difüzyon, Üretici Modeller, Derin Öğrenme, Görüntü Geriçatımı

To my family

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my supervisor, Assoc. Prof. Dr. S. Figen Öktem, for her invaluable guidance, patience, and continuous support throughout my master's studies. Her insights and encouragement have been essential in shaping this thesis, and her mentorship has greatly contributed to my academic and personal development.

I am also thankful to the members of my thesis committee, Assoc. Prof. Dr. Elif Vural and Prof. Dr. Tolga Çukur, for their valuable feedback and constructive criticism. Their expertise and thoughtful suggestions have significantly improved the quality of this work.

This work has been supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) under the 120E505 grant.

I want to extend my heartfelt appreciation to my family. My mother, Hacer, and my brother, Efe Eren, have been my pillars of strength throughout this journey. Their unwavering support, love, and encouragement have been indispensable, and I cannot thank them enough for their belief in me.

Thank you all for your contributions and support. This thesis would not have been possible without you.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ABBREVIATIONS

PR	Phase Retrieval
InDI	Inversion by Direct Denoising
DDRM	Denoising Diffusion Restoration Models
HIO	Hybrid Input-Output
DNN	Deep Neural Networks
TTA	Test Time Augmentation
DFT	Discrete Fourier Transform
SNR	Signal-to-Noise Ratio
GS	Gerchberg-Saxton
ER	Error Reduction
MAP	Maximum A Posteriori
MMSE	Minimum Mean Squared Error
GAN	Generative Adversarial Network
CDP	Coded Diffraction Pattern
ELBO	Evidence Lower Bound
SVD	Singular Value Decomposition
PSNR	Peak Signal-to-Noise Ratio
SSIM	Structural Similarity Index Measure
FID	Fréchet Inception Distance
LPIPS	Learned Perceptual Image Patch Similarity
CLIP	Contrastive Language–Image Pre-training
IQA	Image Quality Assessment
CDI	Coherent Diffractive Imaging

CHAPTER 1

INTRODUCTION

1.1 Phase Retrieval Problems

In its broadest sense, the phase retrieval problem involves reconstructing an unknown signal $\mathbf{x}$ from the measurements expressed as

$$\mathbf{y^2} = |\mathbf{Ax}|^\mathbf{2} + \mathbf{w} \tag{1.1}$$

where $\mathbf{A}$ represents a known linear operator specific to the application, and $\mathbf{w}$ denotes the measurement noise. It is worth mentioning that phase retrieval problems are more difficult to solve than linear inverse problems due to the non-linearity in this forward model. This problem has many important applications in imaging, computer-generated holography, optical computing, crystallography, microscopy, speech processing, optical engineering, and theoretical machine learning, to name a few. Despite their diverse applications and different physical setups, the forward models in phase retrieval converge to a common mathematical framework [1]–[9].

1.2 Fourier Phase Retrieval

The general problem formulation for phase retrieval as given in Eq. 1.1 reduces to the classical Fourier phase retrieval problem when the measurement operator $\mathbf{A}$ is the Discrete Fourier Transform (DFT) matrix. More formally, in the classical phase retrieval problem, the measurements can be modeled as follows:

$$\mathbf{y^2} = |\tilde{\mathbf{F}}\mathbf{x}|^\mathbf{2} + \mathbf{w}, \qquad \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \alpha^2 \text{diag}(|\tilde{\mathbf{F}}\mathbf{x}|^2)) \tag{1.2}$$

Here, $\mathbf{y^2} \in \mathbb{R}^{\sqrt{m}\times\sqrt{m}}$ denotes the noisy Fourier intensity measurements, while $\tilde{\mathbf{F}}$ represents the oversampled discrete Fourier transform (DFT) matrix with dimensions $\sqrt{m} \times \sqrt{m}$. The target image $\mathbf{x} \in \mathbb{R}^{\sqrt{n}\times\sqrt{n}}$ is presumed to be real-valued, non-negative, and of finite support. The term $\mathbf{w} \in \mathbb{R}^{\sqrt{m}\times\sqrt{m}}$ signifies the measurement noise, and α is a scaling factor that adjusts the signal-to-noise ratio (SNR). The noise is generally modeled as Poisson-distributed, but a normal approximation is used in this scenario [10].

DFT of an two-dimensional signal $\mathbf{x} \in \mathbb{R}^{\sqrt{n}\times\sqrt{n}}$ is given by

$$\hat{\mathbf{x}}\left[k_1, k_2\right] = \frac{1}{4\sqrt{n}} \sum_{n_1=0}^{\sqrt{n}-1} \sum_{n_2=0}^{\sqrt{n}-1} \mathbf{x}\left[n_1, n_2\right] e^{-2\pi j \frac{n_1 k_1 + n_2 k_2}{\sqrt{n}}} \tag{1.3}$$

and denoted as $\hat{\mathbf{x}} = \mathbf{F}\mathbf{x}$.

For discrete real-valued signals with finite support in two or more dimensions, the Fourier intensity measurements at discrete frequencies, denoted as $|\mathbf{F}\mathbf{x}|^{\mathbf{2}}$, can uniquely determine the unknown signal $\mathbf{x}$. To ensure uniqueness (aside from trivial ambiguities), for an image with support $\sqrt{n} \times \sqrt{n}$, it is required to provide the magnitude of its $\sqrt{m} \times \sqrt{m}$-point oversampled DFT with $\sqrt{m} \geq 2\sqrt{n} - 1$ [11]. For simplicity, this work sets m to $4n$.

These trivial ambiguities arise from the fact that there are some transformations that do not modify the Fourier magnitude, such as global phase shift, conjugate inversion, and spatial circular shift.

Supposing the Fourier spectrum of $\mathbf{x} \in \mathbb{R}^{\sqrt{n}\times\sqrt{n}}$ is oversampled twice uniformly at $k_i = \{0, 1/2, 1, ..., \sqrt{n} - 1/2\}$ for $i = 1, 2$, we can write this oversampled spectrum $\hat{\mathbf{x}}^{(2)}$ as

$$\begin{aligned}
\hat{\mathbf{x}}^{(2)}\left[k_1, k_2\right] &= \frac{1}{4\sqrt{n}} \sum_{n_1=0}^{\sqrt{n}-1} \sum_{n_2=0}^{\sqrt{n}-1} \mathbf{x}\left[n_1, n_2\right] e^{-2\pi j \frac{n_1 k_1 + n_2 k_2}{\sqrt{n}}} \\
&= \frac{1}{4\sqrt{n}} \sum_{n_1=0}^{\sqrt{m}-1} \sum_{n_2=0}^{\sqrt{m}-1} \sqrt{\frac{n}{m}} \tilde{\mathbf{x}}\left[n_1, n_2\right] e^{-2\pi j \frac{n_1 \tilde{k}_1 + n_2 \tilde{k}_2}{\sqrt{m}}} \\
&= (\mathbf{F}\tilde{\mathbf{x}})\left[\tilde{k}_1, \tilde{k}_2\right]
\end{aligned} \tag{1.4}$$

where $\tilde{k}_i = \{0, 1, \cdots, 2\sqrt{n} - 1\} = 2k_i$ and $\tilde{\mathbf{x}} \in \mathbb{R}^{\sqrt{m}\times\sqrt{m}}$ with $m = 4n$ is defined such that $\tilde{\mathbf{x}}\left[n_1, n_2\right] = \sqrt{\frac{m}{n}} x\left[n_1, n_2\right]$ if $n_i \in \mathbb{N} < \sqrt{n}$ and $\tilde{x}\left[n_1, n_2\right] = 0$, otherwise.

If the vectorization order is defined such that

$$\tilde{\mathbf{x}}^T = \sqrt{\frac{m}{n}} \begin{bmatrix} \mathbf{x}^T & \mathbf{0}_{m-n}^T \end{bmatrix} \tag{1.5}$$

then

$$\hat{\mathbf{x}}^{(2)} = \mathbf{F}\tilde{\mathbf{x}} = \mathbf{F}\mathbf{O}_{\mathbf{mn}}\mathbf{x} \tag{1.6}$$

where

$$\mathbf{O}_{mn} = \sqrt{\frac{m}{n}} \begin{bmatrix} \mathbf{I}_n \\ \mathbf{0} \end{bmatrix}. \tag{1.7}$$

Defining $\tilde{\mathbf{F}} = \mathbf{F}\mathbf{O}_{\mathbf{mn}}$, the classical phase retrieval problem given in Eq. 1.2 is equivalent to finding a supported signal $\tilde{\mathbf{x}}$ from its noisy DFT intensity measurements with the known support constraint sometimes including the support constraint of $\mathbf{x}$ itself.

It is essential to highlight the critical importance of the Fourier phase in order to understand the difficulty of the problem. The Fourier phase holds more information than the Fourier magnitude, as demonstrated in the example depicted in Fig. 1.1. In this example, two images are transformed to the Fourier domain, their phase components are swapped, and then their inverse Fourier transforms are computed. The resulting images highlight that the Fourier phase retains a substantial amount of the original image information. This experiment demonstrates that the vital importance of phase information which is more significant than magnitude information. In the Fourier phase retrieval problem, we only have access to the Fourier magnitude information, which makes it a challenging task.

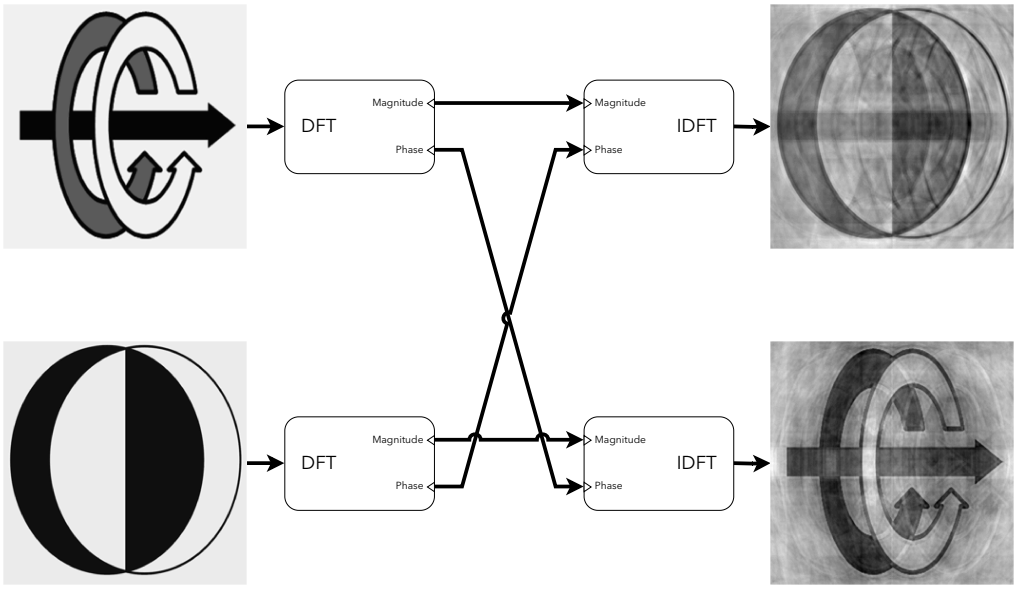


Figure 1.1: Synthetic example showing the importance of Fourier phase information. Two images are Fourier transformed, their phases are swapped, and then inverse Fourier transformed. The resulting images demonstrate that the phase information holds a significant amount of the original image details.

1.3 Its Significance in Optics

In optics, the significance of phase retrieval stems from the inherent limitations of optical detection devices, such as CCD cameras and photosensitive films, which can only measure the intensity of light and not its phase. This limitation is due to the high oscillation frequencies of electromagnetic waves, approximately 10^{15} Hz, which are beyond the capability of electronic devices to capture phase information directly.

Consequently, measuring the phase of optical waves involves additional complexity, typically by requiring interference with another known field through holography/interferometry. In such holographic/interferometric approaches, the phase information exists in amplitude modulation of another wavefront, and the intensity of this wavefront is measured. However, there are situations (e.g. X-ray imaging, imaging through turbulent atmosphere) where setting up an interferometer is not always practical. Therefore, non-interferometric phase retrieval techniques requiring computational algorithms are indispensable when direct phase measurements are impossible.

1.4 An Example Application: Coherent Diffractive Imaging

One of the primary challenges in optical imaging systems is overcoming the diffraction limit, which restricts the resolution to a value proportional to the wavelength of the light used. For visible light, this diffraction limit is in the micron range, making it impossible to image molecular-scale features. Although using electromagnetic waves of shorter wavelengths, such as hard X-rays, could theoretically achieve atomic resolution, practical limitations arise. Many lens-like devices and other optical components suffer from significant aberrations and are difficult to manufacture due to the refractive indices of materials in this spectral region being close to one. Additionally, materials tend to absorb too much of the shorter wavelength light, which further complicates the imaging process and reduces the efficiency and clarity of the resulting images.

This is where coherent diffractive imaging (CDI) comes into play. CDI is a revolutionary lensless imaging technique that has significantly advanced the field of microscopy. It utilizes a coherent light source, such as X-rays or lasers, to illuminate an object and measure the resulting diffraction pattern, which corresponds to the Fourier transform of the object in the far-field. Since only the intensity of the diffraction pattern can be recorded, phase retrieval algorithms are essential for reconstructing the image of the object from these measurements.

The revival of optical phase retrieval in 1999 marked a significant milestone for CDI. Miao et al. successfully recorded and reconstructed a continuous diffraction pattern of a non-crystalline object, demonstrating the potential of phase retrieval to achieve high-resolution imaging without traditional lenses [12]. This breakthrough has allowed CDI to be applied using various sources, including synchrotron radiation, X-ray free electron lasers, high harmonic generation, optical laser, and electrons, enabling high-resolution imaging of non-crystalline samples. The technique has proven particularly transformative in microscopy, offering a powerful tool for imaging small features that are beyond the capabilities of traditional diffraction-limited lens-based systems [1].

In CDI, the object distribution can be described by a complex-valued transmittance

function $T_*(\mathbf{r})$, where $\mathbf{r} = (x, y, z)$ is the coordinate in the object domain. When a plane wave is incident on the object, the distribution of the scattered wave in the detector plane is calculated by the following integral transformation:

$$U(\mathbf{R}) = -\frac{j}{\lambda} \iiint e^{jkz} T_*(\mathbf{r}) \frac{e^{jk|\mathbf{R}-\mathbf{r}|}}{|\mathbf{R}-\mathbf{r}|} \mathrm{d}\mathbf{r}, \tag{1.8}$$

where λ is the wavelength of the employed probing wave, $k = \frac{2\pi}{\lambda}$ is the wavenumber, $\mathbf{R} = (X, Y, Z)$ is the coordinate in the detector plane, e^{jkz} is the incident plane wave, and $|\mathbf{R} - \mathbf{r}|$ is the distance between a point in the object plane and a point in the detector plane. The integration is performed over all scattering elements of the object [13], [14].

Under the Fraunhofer far-field approximation, the model simplifies to the 2D Fourier phase retrieval problem upon sampling if we consider the projected object distribution achieved by integration along the optical axis, $T(x, y) = \int T_*(x, y, z)dz$ [13], [14]. This simplification, illustrated in Fig. 1.2, is a well-known result in Fourier optics [15]. In this context, the observed far-field intensity pattern in the detector plane, $I(x, y)$, is proportional to the Fourier intensity of the unknown transmittance function, $T(x, y)$:

$$I(x, y) \propto \left| \mathcal{F}\{T\}(\frac{x}{\lambda d}, \frac{y}{\lambda d}) \right|^2 \tag{1.9}$$

In this thesis, we focus exclusively on real-valued transmission functions rather than complex-valued. Physically, a sample with a real transmittance function does not cause any phase shift to the incoming coherent wave and can only absorb it.

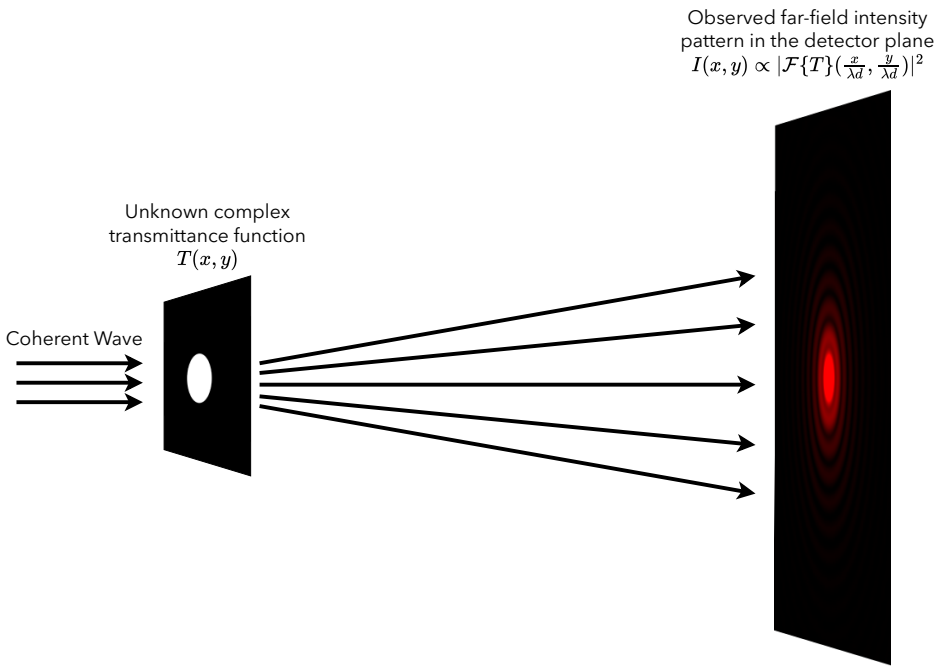


Figure 1.2: Schematic representation of the coherent diffractive imaging (CDI) setup. A coherent wave illuminates an object characterized by an unknown complex transmittance function $T(x, y)$. The resulting diffraction pattern, observed in the far-field detector plane, is proportional to the Fourier intensity of the transmittance function. This observed intensity pattern, $I(x, y)$, is crucial for reconstructing the object's image using phase retrieval algorithms.

1.5 Other Applications and History

It is not difficult to imagine other physical problems that result in the same mathematical phase retrieval formulation. For instance, while CDI focuses on reconstructing the image of an unknown object, we can also consider a synthesis setup, such as in computer-generated holography. In this case, the challenge is to design a transmittance function that produces a specified far-field intensity pattern. This synthesis problem relies on the same phase retrieval process to achieve the intended results.

The historical development of phase retrieval adds further context to its wide-ranging applications. Phase retrieval has a rich history, with its origins tracing back to the mid-20th century in the field of crystallography. In 1952, Sayre proposed that phase

information of a scattered wave could be recovered if the intensity pattern at and between the Bragg peaks was finely measured, leveraging the periodic nature of crystals. This idea was further developed in 1978 when Fienup introduced algorithms for retrieving phases of two-dimensional images from their Fourier modulus using constraints such as non-negativity and known support. These foundational works laid the groundwork for the broader application of phase retrieval techniques across multiple scientific disciplines [1]–[9].

The importance of the phase retrieval problem extends to various other applications, such as astronomy and optical communication. In astronomy, phase retrieval algorithms have been crucial in correcting aberrations in telescope imaging, including the well-known correction of the Hubble Space Telescope's primary mirror aberration. Similar techniques are employed in adaptive optics to compensate for atmospheric distortions, enabling ground-based telescopes to achieve space-quality imaging. In optical communication, phase retrieval is used to design the temporal phase of light beams transmitted through optical fibers, compensating for dispersion and temporally concentrating energy at the output, which is vital for high-speed data transmission [3].

Additionally, phase retrieval plays a key role in beam shaping, where the goal is to design phase-only transparencies to produce desired intensity patterns. This is crucial in applications such as laser machining and inertial confinement fusion. Phase retrieval is also used in optical encryption, where diffractive optical elements with quasi-random phases are designed for secure data reconstruction, and in the iterative design of antireflection coatings and other multilayer optical structures [3].

The application of phase retrieval is not limited to these areas. It also includes wavefront sensing for radio antennas and optics, frequency-resolved optical gating (FROG) for characterizing laser pulses, and tomographic imaging with incomplete projections or unknown phases. These applications demonstrate the broad utility of phase retrieval across various scientific and engineering domains [1]–[9].

In recent years, phase retrieval has gained renewed interest due to the development of new imaging techniques and the integration of modern optimization and machine learning methods. Theoretical and algorithmic advancements have significantly improved the performance and applicability of phase retrieval methods. For instance,

exploiting the sparsity of optical images has led to powerful phase retrieval methods that achieve resolutions beyond the diffraction limit, resolving features smaller than one-fifth of the wavelength [16]. As enforced priors about the unknown signal have become more sophisticated, these methods have been able to provide even greater accuracy and robustness in image reconstruction.

1.6 Classical Iterative Projection Techniques for Phase Retrieval

Iterative projection techniques have become fundamental tools for phase retrieval. One of the earliest and most well-known algorithms is the classical Gerchberg-Saxton (GS) algorithm [17], which iteratively applies magnitude constraints in both the spatial and Fourier domains to reconstruct an unknown signal. An enhancement of the GS algorithm is the Error Reduction (ER) algorithm, which incorporates additional spatial domain constraints beyond just magnitude [18]. A particularly significant and widely used method among alternating projection techniques is the Hybrid Input-Output (HIO) algorithm [19], which builds upon the principles of the ER algorithm.

In the HIO method, Fourier magnitude constraints and various spatial domain constraints (such as support, non-negativity, and real-valuedness) are iteratively applied, similar to the ER algorithm. However, the key distinction is that HIO does not force the iterates to strictly satisfy the constraints at every step. Instead, it uses the iterates to progressively guide the algorithm towards a solution that meets the constraints [19]. The HIO iterations are mathematically expressed as follows:

$$\mathbf{x}_{k+1}[n] = \begin{cases} \mathbf{x}'_k[n] & \text{for} \quad n \notin \gamma \\ \mathbf{x}_k[n] - \beta \mathbf{x}'_k[n] & \text{for} \quad n \in \gamma \end{cases} \tag{1.10}$$

where

$$\mathbf{x}'_k = \mathbf{F}^{-1} \left\{ \mathbf{y} \odot \frac{\mathbf{F}\mathbf{x}_k}{|\mathbf{F}\mathbf{x}_k|} \right\}. \tag{1.11}$$

In these equations, $\mathbf{x}_k \in \mathbb{R}^{\sqrt{m} \times \sqrt{m}}$ represents the reconstruction at the k^{th} iteration, $\mathbf{F^{-1}}$ denotes the inverse Discrete Fourier Transform (DFT) matrix, $\odot$ signifies element-wise multiplication, β is a constant parameter (commonly set to 0.9), and γ is the set of indices n where $\mathbf{x}'_k[n]$ fails to meet the spatial domain constraints [19].

Despite the lack of a comprehensive theoretical understanding of the HIO method's convergence behavior, it has been empirically observed to converge to acceptable solutions in a wide array of applications. However, the reconstructions produced by HIO can sometimes contain artifacts and errors. These issues are often attributed to the algorithm getting trapped in local minima or to the amplification of noise within the solution [1], [20]. To address these limitations, numerous variations and enhancements of the HIO method have been proposed, aiming to improve its reconstruction performance and reliability [21], [22].

1.7 Deep Learning for Inverse Problems

Deep learning-based reconstruction techniques have emerged as a compelling alternative to traditional analytical methods. These approaches demonstrate the potential to achieve high reconstruction quality and computational efficiency across various imaging problems, including phase retrieval [23], [24]. The integration of deep learning into phase retrieval represents a significant advancement, offering new solutions to longstanding challenges. Deep learning priors are particularly useful for phase retrieval because they can effectively capture complex structures and patterns in data, which are difficult to represent with traditional analytical techniques. By learning from large datasets, deep learning models can provide robust priors that guide the phase retrieval process to reduce the impact of noise and improve convergence to accurate solutions [24].

The current landscape of deep learning-based reconstruction in the literature can be broadly categorized into four main classes: 1) learning-based direct inversion, 2) plug-and-play regularization, 3) learned iterative reconstruction based on unrolling, and 4) generative methods.

Learning-based direct inversion methods aim to bypass iterative reconstruction altogether by directly mapping measurements to the desired image using a deep neural network (DNN). This approach trains the DNN to learn the inverse function of the forward model solely on the basis of the training data. While achieving state-of-the-art performance for simpler inverse problems like denoising [25], these methods face

challenges with complex observation models, significant discrepancies between observations and the target image, or limited training data availability. Such end-to-end schemes also exist for the phase retrieval problem [26]–[28]. However, due to the nature of the phase retrieval problem, such end-to-end learning approaches generally do not perform well compared to other approaches [29].

To address these limitations, a common strategy involves applying an efficient analytical approximation of the forward model to generate an initial reconstruction. This initial estimate then serves as a "warm start" for a subsequent DNN refinement step. This hybrid approach, which combines neural networks with analytical methods, has demonstrably succeeded in various real-valued 2D reconstruction problems, including deconvolution, super-resolution, tomography, and phase retrieval [26], [29], [30]. Notably, a key advantage of learning-based direct inversion methods lies in their low computational complexity due to their feed-forward (non-iterative) nature, making them well-suited for real-time imaging applications.

In contrast to learning-based direct inversion, plug-and-play regularization, and unrolled learning methods embrace iterative strategies. Their core principle lies in replacing hand-crafted analytical priors with data-driven deep priors within model-based reconstruction frameworks. Plug-and-play methods first train a deep prior on dedicated datasets and then leverage it as a regularizer for an iterative model-based inversion algorithm [31], [32]. Since maximum a posteriori problem given Gaussian noise assumption can be written as an optimization problem in the form of $\max_{\mathbf{x}} -\|\mathbf{y} - \mathcal{A}(\mathbf{x})\|^2 + \mathcal{R}(\mathbf{x})$ which can be split into data-fidelity and regularization steps, this framework allows to solve various inverse problems by leveraging the impressive capabilities of existing denoising models in the regularization steps while model-based algorithms can be used jointly in the data-fidelity steps. Such plug-and-play methods are widely used in the current phase retrieval literature [33]–[36]. While achieving superior image quality, flexibility, and generalizability compared to direct inversion methods, these approaches typically incur higher memory usage and computational complexity due to their iterative nature. This complexity stems from the need to compute the forward operator (system model) and its adjoint at each iteration.

Unrolled learning takes iterative methods utilizing proximal operators or deep pri-

ors, such as those employed in plug-and-play approaches, and transforms them into end-to-end trainable networks. This representation allows the algorithm to be concatenated as a series of layers, running a finite number of times as it passes through the network. This unrolling aims to further improve reconstruction quality [37], [38]. However, similar to plug-and-play methods, unrolled iterative learning generally suffers from high computational demands. Furthermore, unlike direct inversion and plug-and-play methods, unrolled approaches necessitate the computation of both forward and adjoint operators during training, leading to a significant increase in training time and complexity. This can make them impractical for large-scale reconstruction problems. Despite these limitations, unrolled learning has shown success in phase retrieval [39]–[43].

1.8 Generative Models for Inverse Problems

All of the aforementioned deep learning methods focus on Maximum A Posteriori (MAP) or Minimum Mean Squared Error (MMSE) estimation. As theoretically shown in [44] and empirically observed in [45], these estimates may deviate significantly from the natural image manifold, leading to reconstructions with overly smooth features. Interestingly, the work by Işıl et al. [33] attributes this smoothing behavior to an unavoidable inherent limitation of Deep Neural Networks (DNNs) in the context of phase retrieval. However, as long as reconstruction algorithms prioritize minimizing metrics like MSE, we can only expect limited improvements in perceptual quality.

To achieve reconstructions that are visually accurate to human observers, a shift in our strategy for solving inverse problems is necessary. Instead of focusing solely on the conditional mean of the posterior distribution, we should aim to sample directly from this posterior distribution $p(\mathbf{x}|\mathbf{y})$. This allows us to generate images that are more likely to belong to the true underlying distribution of natural images.

In cases of severe information loss, the image reconstruction problem becomes ill-posed, meaning that there can be multiple valid solutions (clean images) that explain the observed measurements. This challenge is particularly relevant in phase retrieval,

where intrinsic system symmetries can map different input images to the same output, which affects network performance [46]. The MMSE solution attempts to average these potential solutions, resulting in smoothed images lacking the fine details often present in real-world scenes. Given the existence of multiple valid solutions, a successful approach should incorporate stochasticity, as ill-posed problems inherently have multiple viable solutions for the same data. Generative models provide an ideal framework for this purpose, allowing us to sample from the posterior distribution and generate diverse yet plausible reconstructions.

Generative models, which include techniques such as Generative Adversarial Networks, Variational Autoencoders, flow-based approaches, and diffusion models, have demonstrated impressive performance in diverse inverse problem tasks [47], [48]. By learning to generate samples from the posterior distribution, generative models can produce reconstructions that better capture the variability and richness of natural images. Notably, generative models have also been successfully applied to phase retrieval [27], [49], [50]. Uelwer et al. [27] demonstrated that conditional generative adversarial networks (cGANs) can optimize phase retrieval processes by incorporating measurement knowledge, thus achieving superior performance compared to traditional methods. Similarly, Gladrow et al. [49] utilized deep conditional generative models like cGAN and cVAE to solve the inverse problem of digital holography, showcasing the potential of data-driven approaches in handling optical aberrations. Shoushtari et al. [50] introduced DOLPH, a diffusion model-based architecture, which effectively integrates image priors with nonconvex data-fidelity terms, providing robust and high-quality solutions for phase retrieval. These studies collectively highlight the versatility and robustness of generative models in enhancing phase retrieval outcomes.

1.8.1 Diffusion Models for Inverse Problems

Diffusion models, a subclass of generative models, have recently gained prominence for their effectiveness in high-dimensional data generation and reconstruction tasks. These models work by simulating a diffusion process that transforms simple, noise-like data into complex structures over time. The process is guided by learned score

functions, which estimate the gradients of the data distribution at each step to gradually denoise the data and refine the generated outputs.

The significance of diffusion models lies in their theoretical foundation and practical success. Historically, these models draw inspiration from non-equilibrium thermodynamics and stochastic processes. The seminal works on diffusion models have demonstrated their capability to generate high-quality, diverse samples, rivalling or surpassing other generative models such as GANs and VAEs. The iterative nature of diffusion models allows them to incrementally refine solutions [27], [48]–[51], making them particularly well-suited for tasks requiring high precision, such as phase retrieval.

In the context of phase retrieval, diffusion models provide a powerful framework for incorporating deep learning priors. The iterative denoising process aligns well with the need to progressively refine phase estimates from initial noisy guesses. By training on large image datasets, diffusion models learn to capture the underlying statistical properties of the data, which can then be leveraged to guide the phase retrieval process towards more accurate reconstructions.

One of the key advantages of using diffusion models for phase retrieval is their robustness to noise and initialization. Traditional phase retrieval algorithms often suffer from convergence to local minima and sensitivity to the initial guess. Diffusion models, with their probabilistic and iterative nature, can mitigate these issues by providing a systematic approach to explore the solution space and progressively enhance the quality of the reconstructions.

Moreover, the flexibility of diffusion models allows their adaptation to various types of data and measurement settings. Whether dealing with coded diffraction patterns, multi-plane intensity measurements, or different wavelengths, diffusion models can be trained to incorporate these variations, enabling a unified framework for phase retrieval across diverse applications.

1.9 Proposed Methods and Contributions

In this thesis, three novel phase retrieval methods are developed. In Chapter 2, we present a novel approach to phase retrieval by leveraging Langevin dynamics for posterior sampling. This method diverges from traditional techniques that only prioritize distortion metrics, and instead focuses on the perceptual quality-distortion tradeoff to achieve high-fidelity reconstructions with minimal distortion and high perceptual quality. The chapter introduces two deep learning pipelines, prNet-Small and prNet-Large. These pipelines iteratively refine initial HIO estimates through denoising, data consistency, and noise injection cycles. The prNet-Large model incorporates diverse starting points and an additional denoiser with a Wasserstein loss to enhance robustness and perceptual quality. Moreover, test time augmentation, which exploits the inherent properties of the phase retrieval problem, further enhances the performance of prNet-Large with little additional computational cost. Extensive simulations demonstrate that this method achieves state-of-the-art performance while maintaining low computational overhead. The hybridization of denoisers with model-based techniques as in this approach shows significant promise for developing reliable stochastic solvers for nonlinear inverse problems.

In Chapter 3, we introduce a novel approach to Fourier phase retrieval by employing the Inversion by Direct Denoising (InDI) framework [52]. This methodology features a sophisticated initialization strategy, utilizes ensembling to refine quality metrics, and adapts the InDI process specifically for phase retrieval, to achieve significant improvements in both training efficiency and image quality. By starting from a plausible initial estimate rather than random noise, the method maximizes the capacity of the denoiser, resulting in reduced training time and enhanced performance. This approach demonstrates superior results compared to both classical and recent techniques, highlighting its potential for effective and efficient phase retrieval.

Chapter 4 extends the application of Denoising Diffusion Restoration Models (DDRM) [53] from linear inverse problems to the nonlinear inverse problem of phase retrieval. This chapter combines the efficient, unsupervised posterior sampling method of DDRM with the model-based Hybrid Input-Output (HIO) method. This innovative approach uses pretrained unconditional diffusion models. The efficacy of the proposed method

is evaluated using image quality metrics to compare the ground truth and reconstructed images, demonstrating its potential to outperform existing classical iterative methods in phase retrieval.

We can classify the developed algorithms in terms of reconstruction quality and training time, as illustrated in Fig. 1.3. The figure demonstrates the tradeoffs between the different methods developed in this thesis. The prNet model, presented in Chapter 2, exhibits the highest reconstruction performance but requires the longest training time due to its unrolled-like training strategy. In contrast, the DDRM-PR approach in Chapter 4 requires no training but shows lower reconstruction performance. The InDI-PR method, developed in Chapter 3, offers a balanced approach, providing moderate reconstruction performance with relatively efficient training time. This classification shows the varying strengths and tradeoffs of each method, highlighting the versatility and adaptability of the proposed techniques to different application needs. The InDI-PR method achieves training efficiency through a fixed noise schedule, which simplifies the learning process but restricts the pipeline's adaptability to variable noise conditions in phase retrieval tasks. Unlike the prNet model, which allows for dynamic adjustment of the learnable noise schedule during training, the InDI-PR's predefined forward process limits flexibility, impacting its reconstruction performance and ability to handle complex scenarios.

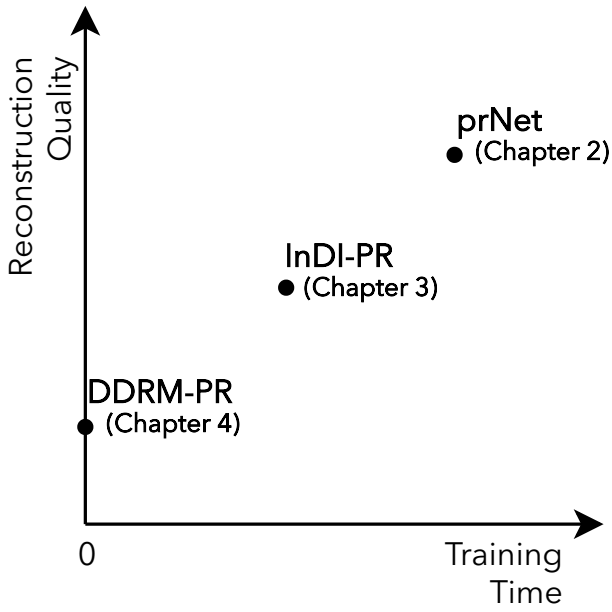


Figure 1.3: Classification of developed phase retrieval algorithms based on reconstruction performance and training time.

1.10 Outline of the Thesis

The rest of this thesis is organized as follows. In Chapter 2, we introduce a novel approach to phase retrieval by leveraging Langevin dynamics for posterior sampling within the framework of diffusion models. We develop two deep learning pipelines, prNet-Small and prNet-Large, which iteratively refine initial estimates to achieve high-fidelity reconstructions with low distortion. Chapter 3 presents the Inversion by Direct Denoising (InDI) framework for Fourier phase retrieval, incorporating advanced initialization strategies and ensembling techniques to enhance training efficiency and image quality. Chapter 4 extends the application of Denoising Diffusion Restoration Models (DDRM) to the nonlinear inverse problem of phase retrieval. We combine generative diffusion models with the Hybrid Input-Output (HIO) method to utilize pretrained unconditional diffusion models for superior phase retrieval performance. Finally, in Chapter 5, we summarize the contributions of this thesis and discuss potential future research directions.

CHAPTER 2

PRNET: SOLVING FOURIER PHASE RETRIEVAL PROBLEM VIA STOCHASTIC REFINEMENT

2.1 Introduction

As mentioned before, Fourier phase retrieval (PR) refers to the problem of reconstructing a signal from the magnitude of its Fourier transform measurements. It arises in numerous applications across science and engineering, including crystallography, microscopy, astronomy, optical imaging, and speech processing [1]–[9]. Despite its widespread utility, PR remains a challenging ill-posed inverse problem due to the loss of phase information and the associated non-convexity. Over the years, numerous algorithmic approaches have been proposed to tackle this problem, each with its own strengths and limitations.

Classical methods for PR typically employ alternating projection schemes that iterate between enforcing the known magnitude constraints in the Fourier domain and imposing prior signal constraints in the spatial domain. A prominent example is the Hybrid Input-Output (HIO) algorithm, which benefits from computational efficiency but may yield suboptimal reconstructions due to stagnation in local minima or noise amplification. More advanced techniques based on semidefinite programming, sparse regularization, and global optimization have also been developed to mitigate these drawbacks, albeit at increased computational cost or restrictive assumptions.

In recent years, deep learning has emerged as a powerful tool for solving various inverse problems in imaging, including PR. Data-driven approaches based on deep neural networks (DNNs) have demonstrated remarkable success in directly reconstructing images from measurements or refining initial estimates from classical meth-

ods. Alternatively, model-based optimization schemes have been augmented with deep priors learned from data using the plug-and-play framework. However, existing deep learning solutions for PR often suffer from limited performance due to domain shift, which occurs when the training data and the real-world test data come from different distributions, leading to decreased accuracy. Additionally, these solutions face challenges related to a lack of interpretability and the need for cumbersome parameter tuning.

Despite extensive research on PR for coded diffraction patterns, a notable research gap exists in the context of classical Fourier PR, with a scarcity of dedicated literature and limited investigations in this setting. Thus, the developed method in this chapter is important as it focuses specifically on this less-explored area.

In this work, we present a novel hybrid approach that synergistically combines model-driven and data-driven techniques for Fourier PR. Our main contributions are as follows:

- We develop two new deep learning pipelines, prNet-Small and prNet-Large, that achieve state-of-the-art reconstruction performance while maintaining low computational time.
- Our methods integrate model-driven and data-driven approaches through iterative refinement of initial HIO estimates using denoising, data consistency, and noise injection cycles guided by deep neural networks.
- prNet-Large incorporates diverse starting points and employs an extra denoiser with a Wasserstein loss to enhance robustness and perceptual quality.
- We introduce the first method that leverages test time augmentation (TTA) for enhanced image reconstruction in Fourier PR, capitalizing on the inherent properties of the problem.

Through extensive simulations, our proposed methods show superior performance compared to classical and state-of-the-art techniques, underscoring their efficacy in addressing the challenges inherent to PR. Moreover, the hybridization of denoisers

with model-based approaches demonstrates promise for developing reliable stochastic nonlinear inverse problem solvers, which could have broader implications beyond PR.

The subsequent sections of this chapter unfold as follows: Section 2.2 reviews related research that informed the development of our approach. Our developed approach is detailed in Section 2.3, followed by a comparative performance analysis against classical and state-of-the-art methods in Section 2.4. Lastly, Section 2.5 summarizes our findings and outlines future research directions.

2.2 Related Works

2.2.1 Posterior Sampling via Score/Diffusion-Based Models

Unconditional diffusion/score-based models are known for their ability to generate high-quality samples from a prior distribution using the score function $\nabla_x \log p(\mathbf{x})$ via Langevin dynamics. It is worth mentioning that since score-based and diffusion-based interpretations are equivalent thanks to Tweedie's formula [51], [54], [55], we can focus solely on the score-based approach here.

Although directly learning the score function is an option, most work utilizes a deep denoiser instead. This substitution is based on the relationship given by [56]

$$\nabla_{\mathbf{x}_t} \log p\left(\mathbf{x}_t\right) = \frac{\text{Denoiser}\left(\mathbf{x}_t, \sigma\right) - \mathbf{x}_t}{\sigma_t^2} \tag{2.1}$$

where $\mathbf{x}_t = \mathbf{x} + \mathbf{v}$ with $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma_t \mathbf{I})$.

Several strategies have been explored to extend this score-based approach for sampling from a posterior distribution $p(\mathbf{x}|\mathbf{y})$, leveraging the posterior score function $\nabla_x \log p(\mathbf{x}|\mathbf{y})$ within Langevin dynamics. Here, we can discuss four common methods for approximating the posterior score function for inverse problems: 1) conditioning via initialization, 2) conditional denoiser, 3) hard projection, and 4) Bayesian approach.

The "conditioning via initialization" approach initializes Langevin dynamics with a plausible estimate obtained from a simpler method, but it does not modify the score

function itself. While this method is simple to implement, it lacks a guarantee of consistency with the observations. Consequently, the resulting outputs may be inconsistent with the actual measurements.

In "conditional denoiser" techniques, they give $\mathbf{y}$ to denoiser as $\nabla_{\mathbf{x}_t} \log p\left(\mathbf{x}_t|\mathbf{y}\right) = \frac{\text{Denoiser}(\mathbf{x}_t,\mathbf{y},\sigma)-\mathbf{x}_t}{\sigma_t^2}$ and rely fully on the learning process. This approach is still simple, but for many inverse problems, the influence of the estimation can be very challenging to learn as it requires learning the complex measurement model. Also, there is no theoretical guarantee for conditioning.

The "hard projection" methods utilize a regular denoiser followed by a projection step to match with $\mathbf{y}$, more mathematically, $\hat{\mathbf{x}} = \arg\min_{\mathbf{z}} \frac{1}{2}\|\mathbf{z}-\text{Denoiser}(\mathbf{x},\sigma)\|^2$ s.t. $\mathbf{y} = \mathcal{A}(\mathbf{z})$. Although relatively simple to implement, this approach might not be applicable to all inverse problems. Additionally, it can suffer from inaccuracies as the projection step might not achieve perfect conditioning on the measurement.

The "Bayesian" approaches leverage Bayes' rule to derive the posterior score function as $\nabla_{\mathbf{x}} \log p\left(\mathbf{x}_{t-1}|\mathbf{y}\right) = \nabla_{\mathbf{x}} \log p\left(\mathbf{y}|\mathbf{x}_{t-1}\right) + \nabla_{\mathbf{x}} \log p\left(\mathbf{x}_{t-1}\right)$. Offering a mathematically well-founded approach for posterior sampling, this method has been successfully applied to linear inverse problems by Kawar et al. [57].

Therefore, one promising approach for achieving high perceptual quality reconstructions is to employ a posterior score-based sampler, as demonstrated by Kawar et al. [57]. This strategy offers a multitude of potential solutions for attaining perfect perceptual quality, albeit potentially at the expense of PSNR (Peak Signal-to-Noise Ratio) metrics.

2.2.2 Wassertein Adversarial Loss

Generative Adversarial Networks (GANs) have demonstrated remarkable success in creating realistic images. GANs can be employed to tackle inverse problems while producing high-quality outputs [58]. These solvers strive to generate a diverse array of images that not only match the given measurements but also align with the distribution of clean examples. In the realm of phase retrieval (PR), several methods already utilize GANs [27], [59]. However, a significant drawback of GAN-based ap-

proaches for inverse problems is their tendency to assume noiseless measurements, as highlighted in [60], a condition that is rarely encountered in practical scenarios.

Instead of solely depending on GANs, adversarial loss can also be used to remedy over-smoothing due to the optimization of distortion metrics by defining the training loss in the following way [44]:

$$\ell_{\text{total}} = \ell_{\text{distortion}} + \lambda \ell_{\text{adv}} , \tag{2.2}$$

where the first term is the distortion between the original and reconstructed images, and the second term is the standard GAN adversarial loss.

Although the standard GAN adversarial loss has been widely used in various applications, it suffers from several drawbacks that can hinder optimization and affect the quality of generated samples. One significant issue is its reliance on the Jensen-Shannon (JS) divergence metric, which can be problematic when the distributions of real and generated samples fall apart. This can lead to training instability, mode collapse, and poor sample quality. In contrast, the Wasserstein GAN (WGAN) introduces a more stable and meaningful metric based on the Wasserstein distance, also known as the Earth Mover's distance, which provides a smoother and more reliable measure of the dissimilarity between distributions. By minimizing the Wasserstein distance, WGAN encourages the generator to produce samples that gradually transition towards the distribution of real data, leading to more stable training dynamics and improved sample quality [61].

An improved version of WGAN, known as WGAN with Gradient Penalty (WGAN-GP), further enhances the training stability and sample quality by imposing a gradient penalty on the discriminator. This penalty term penalizes the norm of the gradients of the discriminator with respect to its inputs, thereby enforcing the Lipschitz continuity. By constraining the Lipschitz constant of the discriminator, WGAN-GP mitigates the risk of mode collapse and training instability while promoting smoother convergence. The algorithm for WGAN-GP involves alternately updating the discriminator and generator parameters while incorporating the gradient penalty term into the loss function. This regularization technique not only improves the robustness of the discriminator/critic but also facilitates a more effective training of the generator, resulting in higher-quality generated samples [62]. Overall, by leveraging the Wasserstein

distance and integrating gradient penalty regularization, WGAN-GP offers a more reliable and effective framework for training GANs, particularly in applications such as inverse problems where stability and sample quality are paramount.

2.2.3 Test Time Augmentation

Test time augmentation (TTA) is a powerful technique in deep learning that leverages data properties to enhance performance without an additional training requirement. It involves creating slightly modified versions of the test images (flips, rotations, crops) and feeding them through the trained model. The predictions from these augmented versions are then combined (typically by averaging) to produce a final prediction [63]. This approach acts as a form of ensembling, effectively increasing the training data by leveraging the inherent equivariance properties of the model and the data distribution [64].

TTA is particularly beneficial when models struggle with small input variations. In image classification, for instance, flipping an image might not significantly alter the content, but the model could potentially misclassify the flipped version. By combining predictions from both versions, TTA achieves a more robust and generalizable performance. This strategy has demonstrably improved accuracy and robustness across various deep learning domains, including image classification [65], object detection [66], and image segmentation [67].

In image reconstruction tasks like ours, TTA can capitalize on the algorithm's invariances. By processing different versions of the test image, TTA integrates additional information at test time, ultimately enhancing the reconstructed output quality. It also mitigates the model's vulnerability to spatial transformations and noise patterns in test data that might have been underrepresented or absent during training.

2.3 Developed Methods

The core of our method is the Langevin dynamics algorithm, which can be used to generate samples from a given posterior probability distribution, $p\left(\mathbf{x}_{t-1}|\mathbf{y}\right)$, based on

the score function of the posterior distribution and a noise term. This is represented by the following equation:

$$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} + \tilde{\alpha}\nabla_{\mathbf{x}} \log p\left(\mathbf{x}_{t-1}|\mathbf{y}\right) + \sqrt{2\tilde{\alpha}}\mathbf{v}_t, \quad 1 \leq t \leq T \tag{2.3}$$

Here, $\tilde{\alpha}$ is the learning rate, $\mathbf{v}_t$ represents Gaussian noise, and T is the number of iterations. Applying Bayes' rule yields:

$$\nabla_{\mathbf{x}} \log p\left(\mathbf{x}_{t-1}|\mathbf{y}\right) = \nabla_{\mathbf{x}} \log p\left(\mathbf{y}|\mathbf{x}_{t-1}\right) + \nabla_{\mathbf{x}} \log p\left(\mathbf{x}_{t-1}\right) \tag{2.4}$$

We can use a trained denoiser to approximate the score function under the assumption of $\mathbf{x}_t = \mathbf{x} + \mathbf{v}$ where $v \sim \mathcal{N}(\mathbf{0}, \sigma_t\mathbf{I})$ [56].

$$\nabla_{\mathbf{x}_t} \log p\left(\mathbf{x}_t\right) = \frac{\text{Denoiser}\left(\mathbf{x}_t, \sigma_t\right) - \mathbf{x}_t}{\sigma_t^2} \tag{2.5}$$

Substituting this into the Langevin dynamics equation, appropriately choosing constants yields:

$$\mathbf{x}_t \leftarrow \text{Denoiser}(\mathbf{x}_{t-1}, t-1) + \tilde{\alpha}\nabla_{\mathbf{x}} \log p\left(\mathbf{y}|\mathbf{x}_{t-1}\right) + \sqrt{2\tilde{\alpha}}\mathbf{v}_t, \quad 1 \leq t \leq T \tag{2.6}$$

As we later make all these terms learnable, the scaling does not affect anything, as they can learn accordingly.

For successful conditioning of the generation process on the observation, the formulation of the likelihood-related term specific to the inverse problem is crucial. Considering the measurement model:

$$\mathbf{y^2} = |\tilde{\mathbf{F}}\mathbf{x}|^\mathbf{2} + \mathbf{w}, \qquad \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \alpha^2 \text{diag}(|\tilde{\mathbf{F}}\mathbf{x}|^2)) \tag{2.7}$$

Under the assumption of $\frac{|\tilde{\mathbf{F}}\mathbf{x}|^\mathbf{2}}{\alpha|\tilde{\mathbf{F}}\mathbf{x}|} >> \mathbb{1}$, the model can be approximated as

$$\mathbf{y} \sim p(\mathbf{y}) \propto \mathbf{y} \odot \mathcal{N}(\mathbf{y^2}; |\tilde{\mathbf{F}}\mathbf{x}|^\mathbf{2}, \alpha^2 \text{diag}(|\tilde{\mathbf{F}}\mathbf{x}|^2)) \approx \mathcal{N}(\mathbf{y^2}; |\tilde{\mathbf{F}}\mathbf{x}|^\mathbf{2}, \alpha^2 \text{diag}(|\tilde{\mathbf{F}}\mathbf{x}|^2)) \tag{2.8}$$

Additionally, by considering the form of a normal distribution:

$$\mathcal{N}(x^2; \mu, \sigma^2) \propto e^{-\frac{(x^2-\mu)^2}{2\sigma^2}} = e^{-\frac{(x-\sqrt{\mu})^2(x+\sqrt{\mu})^2}{2\sigma^2}} = e^{-\frac{(x-\sqrt{\mu})^2}{2(\frac{\sigma}{x+\sqrt{\mu}})^2}} \approx e^{-\frac{(x-\sqrt{\mu})^2}{2(\frac{\sigma}{2\sqrt{\mu}})^2}} \tag{2.9}$$

We can approximate the measurement model as follows:

$$\mathbf{y} = |\tilde{\mathbf{F}}\mathbf{x}| + \mathbf{w}, \qquad \mathbf{w} \sim \mathcal{N}(\mathbf{0}, (\frac{\alpha}{2})^2 \text{diag}(|\tilde{\mathbf{F}}\mathbf{x}|)) \tag{2.10}$$

After assuming independence between $\mathbf{y}$ and $\mathbf{x}_t$, the gradient of the log-likelihood term with respect to $\mathbf{x}_t$ becomes:

$$\nabla_{\mathbf{x}} \log p\left(\mathbf{y}|\mathbf{x}_t\right) \propto \nabla_{\mathbf{x}} \|\mathbf{y} - |\tilde{\mathbf{F}}\mathbf{x}_t|\|^2 \tag{2.11}$$

Also, considering the one-dimensional case without loss of generality, the Wirtinger derivative for a function $\tilde{\mathbf{x}}_n$ with a known support $\mathbf{s}_n$ is expressed as:

$$\begin{aligned}\frac{\partial}{\partial \tilde{\mathbf{x}}_n^*} \sum_i (\mathbf{y}_i - |\mathbf{F}\tilde{\mathbf{x}}|_i)^2 &\propto -\sum_i (\mathbf{y}_i - |\mathbf{F}\tilde{\mathbf{x}}|_i) \frac{\partial |\mathbf{F}\tilde{\mathbf{x}}|_i}{\partial \tilde{\mathbf{x}}_n^*} = -\sum_i (\mathbf{y}_i - |\mathbf{F}\tilde{\mathbf{x}}|_i) \frac{\partial |\mathbf{F}\tilde{\mathbf{x}}|_i}{\partial (\mathbf{F}\tilde{\mathbf{x}})_i^*} \frac{\partial (\mathbf{F}\tilde{\mathbf{x}})_i^*}{\partial \tilde{\mathbf{x}}_n^*} \\ &\propto \sum_i (|\mathbf{F}\tilde{\mathbf{x}}|_i - \mathbf{y}_i) \frac{(\mathbf{F}\tilde{\mathbf{x}})_i}{|\mathbf{F}\tilde{\mathbf{x}}|_i} e^{2\pi j \frac{in}{\sqrt{m}}} s_n \end{aligned} \tag{2.12}$$

Therefore, this leads to the gradient expression:

$$\nabla_{\tilde{\mathbf{x}}} \log p\left(\mathbf{y}|\tilde{\mathbf{x}}_t\right) \propto \text{diag}(\mathbf{s}) \left(\tilde{\mathbf{x}}_t - \mathbf{F}^{-1} \left(\frac{\mathbf{F}\tilde{\mathbf{x}}_t}{|\mathbf{F}\tilde{\mathbf{x}}_t|} \odot \mathbf{y} \right) \right) \tag{2.13}$$

Employing gradient lookahead, where $\tilde{\mathbf{x}}_t = \mathbf{O}_{mn}\text{Denoiser}(\mathbf{x}_t, t)$, involves using the gradient after the denoising step rather than the gradient at $\mathbf{x}_t$. Thus, the Langevin dynamics update is given by:

$$\begin{aligned}\tilde{\mathbf{x}}_t - \tilde{\alpha} \nabla_{\mathbf{x}} \log p\left(\mathbf{y}|\tilde{\mathbf{x}}_t\right) &= \text{diag}(\mathbf{s}) \left((1-\lambda)\tilde{\mathbf{x}}_t + \lambda \mathbf{F}^{-1} \left(\frac{\mathbf{F}\tilde{\mathbf{x}}_t}{|\mathbf{F}\tilde{\mathbf{x}}_t|} \odot \mathbf{y} \right) \right) \\ &= \text{diag}(\mathbf{s}) \mathbf{F}^{-1} \left(\frac{\mathbf{F}\tilde{\mathbf{x}}_t}{|\mathbf{F}\tilde{\mathbf{x}}_t|} \odot (\lambda \mathbf{y} + (1-\lambda)|\mathbf{F}\tilde{\mathbf{x}}_t|) \right) \\ &= \text{ER}(\lambda \mathbf{y} + (1-\lambda)|\mathbf{F}\tilde{\mathbf{x}}_t|) \end{aligned} \tag{2.14}$$

where $\lambda = \tilde{\alpha}$ is a scalar that controls the extent of the measurement update.

Thus, we arrive at one iteration of the Error Reduction (ER) algorithm. However, (sub)gradient methods are known to perform suboptimally for the phase retrieval problem. To address this, we can substitute this step with the Hybrid Input-Output (HIO) algorithm, which demonstrates better convergence properties in practice. Previous diffusion model methods for PR, such as those outlined in [50], did not consider this fact.

It is worth noting that the gradient $\nabla_{\mathbf{x}} \log p\left(\mathbf{y}|\mathbf{x}_t\right)$ was incomplete because we also have additional prior information about $\mathbf{x}$ beyond its support, such as its realness and positiveness. To simulate the effect of $\nabla_{\mathbf{x}} \log p\left(\mathbf{y}, \text{additional constraints}|\mathbf{x}_t\right)$, rather than hard enforcement, we can enforce such constraints in the HIO feedback steps to improve the convergence:

$$\mathbf{x}_t \leftarrow \text{HIO}(\lambda_t \mathbf{y} + (1-\lambda_t)|\tilde{\mathbf{F}}\text{Denoiser}(\mathbf{x}_{t-1}, t-1)|) + \sqrt{2\tilde{\alpha}}\mathbf{v}_t, \quad 1 \leq t \leq T \tag{2.15}$$

Here, we also make the value of λ learnable and time-dependent to optimize the measurement update during the training process.

It is worth mentioning that since the regularization term in the objective function of plug-and-play methods can be viewed as $\log p(\mathbf{x}|\mathbf{y})$, applying the subgradient method to this term reveals a clear connection between PnP and score-based methods. Consequently, it is unsurprising to find PnP methods in the literature that employ iterations strikingly similar to the iteration we derived [34]. But, our work is quite different from it, as they do not target sampling from the posterior.

Also, even though we formulated and tested our method for Fourier magnitude measurements, it can also be utilized for CDP measurements.

The proposed pipeline outlined in Algorithm 1 and Fig. 2.1 employs an efficient utilization of the denoiser's model capacity. To optimize conditioning on observations, we initiate the process with a warm start procedure detailed in [10]. By starting with a plausible estimate rather than a complete noise image, we facilitate training since it should only learn to correct this initial estimate. Note that classical methods, such as HIO, can already provide a decent solution to the problem. Thus, it is logical to start with that decent solution instead of starting with a random noise image so that the denoiser model's capacity is not wasted for the initial reconstruction steps. This "image-to-image" instead of "noise-to-image" idea is also popular in the literature [68]–[71]. However, for such a scheme, we should diverge from the classical diffusion sampling procedure, starting with a pure noise image.

Due to the inherent nonlinearity and non-convexity of the phase retrieval problem, reconstruction algorithms are highly susceptible to the initial guess. In order to address this challenge and enhance the robustness of our method, this initialization procedure

Algorithm 1 Proposed algorithm: prNet-Small

Input: $\mathbf{y}$, $T, K, \tilde{\alpha}, \beta, \lambda \in \mathbb{R}^T$ is learnable (initially, a logarithmically decreasing vector)

Output: $\mathbf{z}_T^{(0)}$

Initialization:

$\mathbf{x'_0} \leftarrow$ HIO initialization procedure

Main loop:

for $i = 1$ to T **do**

  $\mathbf{x}_i \leftarrow$Denoiser($\mathbf{x}'_{i-1}$, i)

  if $i = T$ **then**

    return $\mathbf{x}_i$

  $\mathbf{z}_i^{(0)} \leftarrow \mathbf{O_{mn}}\mathbf{x}_i$

  $\mathbf{y}_i{}' \leftarrow \lambda_i \mathbf{y} + (1 - \lambda_i)|\mathbf{F}\mathbf{z}_i^{(0)}|$

  for $k = 1$ to K **do**

    $\mathbf{z}_i^{(k)'} \leftarrow \Re\{\mathbf{F^{-1}}[\mathbf{y}_i{}' \odot \frac{\mathbf{F}\mathbf{z}_i^{(k-1)}}{|\mathbf{F}\mathbf{z}_i^{(k-1)}|}]\}$

    $\gamma \leftarrow$ the set of indices where $\mathbf{z}_i^{(k)'}$ violates space domain constraints (e.g., support and non-negativity)

    $$\mathbf{z}_i^{(k)}[n] \leftarrow \begin{cases} \mathbf{z}_i^{(k)'}[n] & , n \notin \gamma \\ \mathbf{z}_i^{(k-1)}[n] - \beta \mathbf{z}_i^{(k)'}[n] & , n \in \gamma \end{cases}$$

  $\epsilon \leftarrow \mathcal{N}(\mathbf{0}, \mathbf{I})$

  $\mathbf{x}'_i \leftarrow \sqrt{\frac{n}{m}} \mathbf{O_{mn}^T} \mathbf{z}_i^{(K)} + \tilde{\alpha}\lambda_i \epsilon$

runs the HIO algorithm for a small number of s iterations for m different random phase initializations. This initial exploration aims to identify promising regions in the search space and is highly parallelizable. After selecting the reconstruction with the lowest residual $\|\mathbf{y} - |\mathbf{F}\hat{\mathbf{x}}|\|_2^2$, this reconstruction is then further refined using HIO for a larger number of n iterations.

After the initialization stage, in the main loop, we iteratively apply denoising and data consistency blocks followed by noise addition to get the final reconstruction. Despite HIO's advantages, noise and local minima can still introduce artifacts in reconstructions. To address this, we employ an iterative denoising-data consistency approach also used in many unrolling methods [37]. This scheme aims to escape local minima and reduce artifacts, leading to improved results.

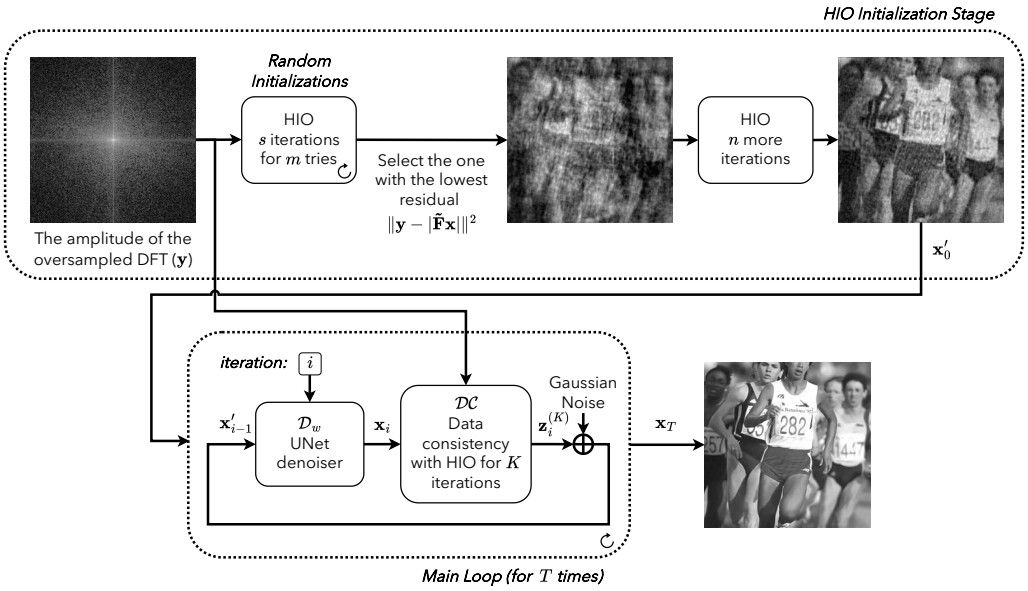


Figure 2.1: The overall pipeline of prNet-Small.

The prNet-Large pipeline given in Fig. 2.2 involving multiple reconstructions enhances overall reconstruction quality. Specifically, the HIO initialization procedure generates k distinct outputs, each of which undergoes denoising, yielding $k/2$ estimations. Following the data consistency layer, the outputs of both the data consistency and denoising stages are concatenated, introducing Gaussian noise before the subsequent iteration. Thus, the denoiser always takes k inputs and generates k outputs.

Note that once we have a systematic way of sampling from $p(\mathbf{x}|\mathbf{y})$, we can also find the MMSE solution by averaging multiple outputs to make the distortion metrics better as MMSE solution gives the best distortion metrics. That explains why ensembling used in the main loop of prNet-Large can make the distortion metrics better.

Following this main loop, a final denoiser refines the output more in order to compensate for the decrease in perceptual quality introduced by ensembling. Our training loss includes an additional Wasserstein GAN loss term, augmenting perceptual quality considerations alongside distortion metrics. This supplementary term serves to balance the distortion-perception tradeoff. Smooth outputs can be generated by a denoiser trained purely based on distortion-related loss metrics such as MSE. But, since a critic model can easily discriminate such smooth outputs, this extra loss term discourages such reconstructions. This Wasserstein GAN loss is directly related to the perceptual quality of the generated samples. As both Langevin dynamics and this added term explicitly address the perception-distortion tradeoff, our algorithm holistically considers this tradeoff as in [44].

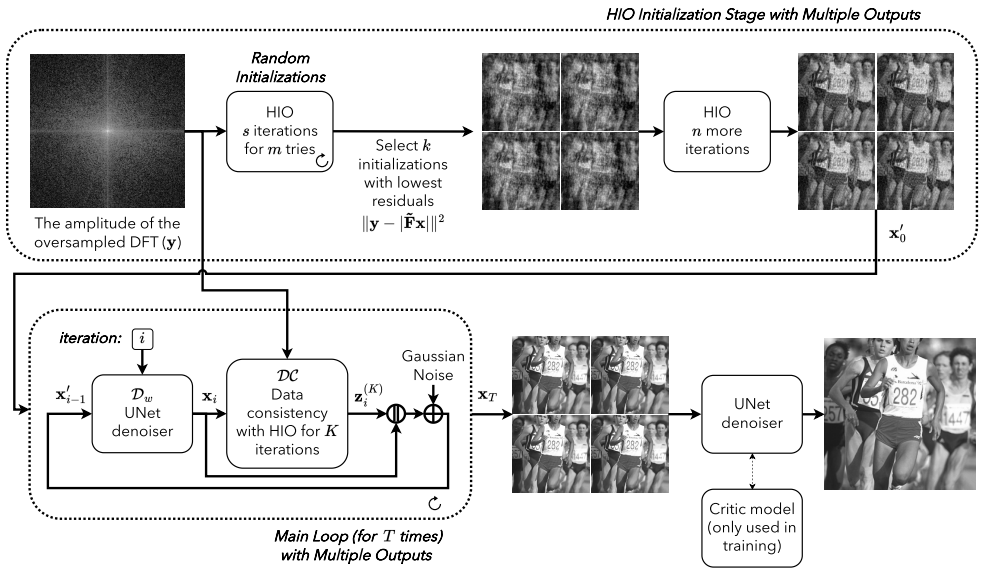


Figure 2.2: The overall pipeline of prNet-Large.

As the denoising component of our pipeline, we employed a customized UNet architecture depicted in Fig. 2.3, a well-established framework renowned for its efficacy in image restoration tasks. Notably, this implementation of UNet incorporates timestep

information as an additional input that is intricately linked to the noise level of the input image. It is imperative to underscore that the denoiser operates by estimating the residual, thus facilitating the refinement of the reconstructed image by focusing only on the discrepancy between the noisy input and the desired clean output. Our customization of the UNet architecture includes blocks that utilize attention mechanisms. These mechanisms enable the network to selectively focus on relevant parts of the input image, enhancing its ability to capture intricate details and effectively suppress noise. This incorporation of attention mechanisms is crucial for improving denoising performance, particularly in scenarios where noise levels vary across different regions of the image.

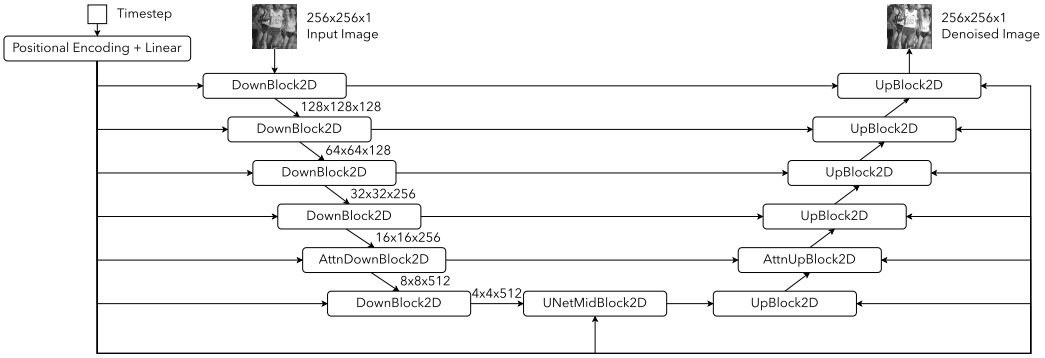


Figure 2.3: Architecture of the UNet denoiser with timestep input.

As in Fig. 2.4, for the training of our pipeline, we adopt a progressive approach that evolves throughout epochs to effectively train each iteration. Initially, we prioritize training the initial iterations to refine the early stages of reconstruction. Within each epoch, we gradually increase the mean of the random timesteps used for training, facilitating a nuanced learning process that adapts to growing temporal complexities. As we approach later epochs, our focus shifts towards training the final iterations. This strategic progression reflects the underlying principle of utilizing outputs from preceding iterations to train subsequent ones, akin to unrolled algorithms, thereby enhancing training efficacy and coherence. In contrast to pipelines that assume diffusion-like processes and train denoisers accordingly, our approach relies on exactly utilizing the outputs of previous iterations for training. While the utilization of exact outputs instead of the approximate ones, simplifies learning, it can extend training duration.

One key advantage of our framework lies in its flexibility regarding the denoising schedule. Unlike methods that assume a fixed, pre-defined diffusion process, our pipeline allows for the learning of this schedule during training. This capability allows the model to discover the optimal denoising strategy that best suits the reconstruction task at hand, potentially leading to superior reconstruction quality.

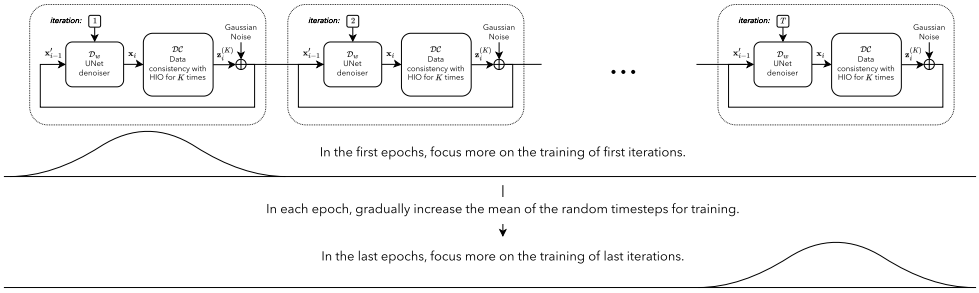


Figure 2.4: Progressive training process.

Since our measurement model is invariant to flipping, meaning that the Fourier magnitude of a flipped image is identical to that of the original, we can leverage this mathematical property for test time augmentation. As depicted in Fig. 2.5, following the robust initialization stage, we can apply flipping to these initialization outputs and execute our pipeline for the flipped versions of the images. Subsequently, combining the flipped outputs with the original outputs allows us to obtain a more refined estimate. Test time augmentation is widely applicable across various deep learning domains and can also be beneficial for enhancing the performance of image reconstruction tasks.

A more advanced Test Time Augmentation technique called TTA D_4, as illustrated in Fig. 2.6, leverages the properties of the D_4 dihedral group, which includes all symmetries of a square, such as rotations and reflections. This method enhances the initial TTA by applying each transformation from the D_4 group to the outputs from the robust initialization stage, covering rotations (R_0, $R_{\pi/2}$, R_π, $R_{3\pi/2}$) and reflections (Horizontal Flip HF, Vertical Flip VF, Diagonal Flip DF, and Anti-Diagonal Flip ADF). Formally, we process the initialization outputs $\{\hat{\mathbf{x}}_{\text{init}}^{(m)}\}_{m=1}^{k}$ with a transform $\mathcal{T}$ to generate new sets of initializations $\{\mathcal{T}(\hat{\mathbf{x}}_{\text{init}}^{(m)})\}_{m=1}^{k}$. We also know the effects of these transformations in the Fourier domain, thus, we also apply the corresponding

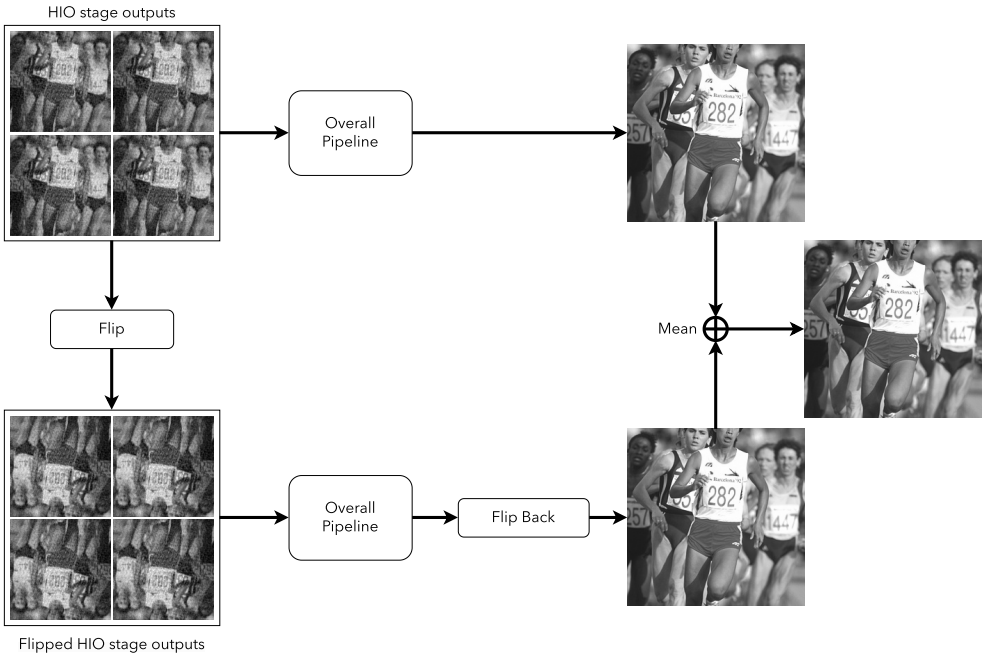


Figure 2.5: Test time augmentation (TTA).

transformation in the Fourier domain to the observation $\mathbf{y}$. These transformed initializations are then iteratively refined, producing different final outputs. The combined final result is obtained by averaging over all D_4 transformations, expressed as:

$$\hat{\mathbf{x}}_{\text{final}}^{(\text{combined})} = \frac{1}{|D_4|} \sum_{\mathcal{T} \in D_4} \mathcal{T}^{-1}(\hat{\mathbf{x}}_{\text{final}}^{\mathcal{T}}) \tag{2.16}$$

where $|D_4| = 8$ is the order of the D_4 dihedral group.

By incorporating all transformations from the D_4 dihedral group, this advanced TTA technique maximizes the use of symmetry properties and available data, significantly enhancing the robustness and quality of image reconstructions. This approach is particularly effective in image reconstruction tasks, where the enriched data from augmentation helps mitigate overfitting and improves generalization performance.

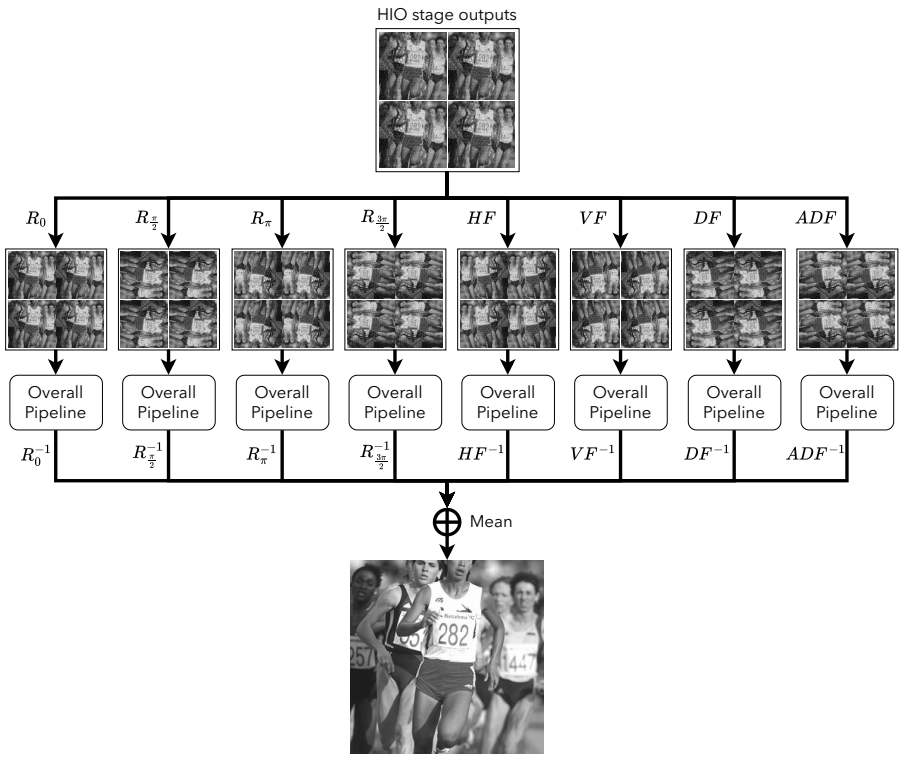


Figure 2.6: Test time augmentation using dihedral group D_4 (TTA D_4).

2.4 Results

To evaluate the performance of our method, we conducted numerical simulations using a large image dataset. We compare the reconstruction performance against both classical and state-of-the-art phase retrieval methods.

To assess the algorithms' robustness to noise, generalization capacity, and computational cost, we investigate their reconstruction performance in two distinct image categories: natural and unnatural. This categorization allows us to evaluate the methods' ability to handle real-world scenarios (natural images) and potentially more challenging, synthetic long-tail samples (unnatural images).

For the training phase, exclusively natural images are utilized. The training dataset comprises $44,000$ natural images, including 200 training and 100 validation images from the Berkeley segmentation dataset (BSD) [72], $41,400$ images selected from the ImageNet database [73], [74], and $2,300$ images randomly chosen from the Waterloo

Exploration Database [75].

During the testing phase, both natural and synthetic images are employed. This test dataset, previously used in [33], [34], contains 236 images, which include 230 natural and 6 synthetic images. Specifically, the dataset consists of 200 test images from BSD, 24 images from the Kodak dataset [76], and 6 natural and 6 synthetic images sourced from [10]. The synthetic image subset features images obtained from scanning electron microscopes and telescopes. All images have pixel values ranging from 0 to 255 and are of size 256×256.

The noisy Fourier measurements were generated using Eq. 1.2, with the average SNR values presented in Table 2.1 (where SNR $= 10\log(\left\||\tilde{\mathbf{F}}\mathbf{x}|^2\right\|_2 / \left\|\mathbf{y^2} - |\tilde{\mathbf{F}}\mathbf{x}|^{\mathbf{2}}\right\|_2)$).

In training, the denoiser model takes the output of the previous iterations as its input and generates an estimate for the clean image. Our training loss includes an MSE loss between this reconstruction and the ground truth image. But, our training loss also has a term for the reconstruction loss of the output of the data consistency block since there are also other learned parameters after the denoising block. Furthermore, for training the final denoiser of the prNet-Large pipeline, an extra improved Wasserstein GAN loss with gradient penalty is added to the training loss. For the critic model used in training, a simple ResNet18 network [77] is used.

Despite the training being conducted solely with natural images, the developed pipeline was evaluated on both natural and synthetic images to assess its generalization capabilities.

Decoupled weight decline regularization [78] is used for optimization together with cosine annealing with linear warming [79]. The developed method is implemented by PyTorch and tested on a NVIDIA A100 80GB PCIe GPU. The total training times for prNet-Small, prNet-Large, and prNet-Large-Adversarial were about four days (for 90 iterations), five days (for 40 epochs), and one day (for 25 epochs), respectively.

In the initialization phase of prNet-Small, the HIO method was initially executed with $m = 50$ different random starting points for $s = 50$ iterations each. The reconstruction with the lowest residual error was selected for an additional HIO run of $n = 1000$ iterations. The resulting reconstruction was then used as input for the

iterative denoiser-HIO stage. In this iterative phase, consisting of $T = 18$ blocks, the HIO method was performed for $K = 5$ iterations before introducing noise under the $\tilde{\alpha} = 9$ setting.

The selected hyperparameters for the prNet-Large pipeline differ from the prNet-Small pipeline only in the initialization stage. In the prNet-Large initialization stage, $k = 10$ multiple outputs are generated from the best $k = 10$ initializations with the lowest residuals among the $m = 100$ different random initializations.

After the testing phase, the reconstructions of the developed approach were compared with the true images using the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [80]. For comparison, results for the same test set are also included for prDeep [10], classical HIO algorithm [18], DIR [33], and MBwDDP [34].

Table 2.1 presents the average reconstruction performance of the algorithms for 236 test images over 5 Monte Carlo runs under varying levels of Poisson noise ($\alpha = 2, 3, 4$). As demonstrated in the table, the developed methods consistently surpass other methods in both PSNR and SSIM metrics across all noise levels, while only necessitating a marginal increase in runtime compared to the robust HIO-based initialization procedure. As another benchmark, the results at the output of different stages of developed algorithms are also provided in the table to show performance gains. The superiority of our methods can also be seen visually in Figs. 2.7 and 2.8. More example reconstructions for test images can be seen in Appendix A.

The results illustrate that, with the prNet-Small pipeline, HIO artifacts can be successfully removed while preserving the image characteristics. prNet-Large with TTA provides the best reconstruction performance since it considers different initializations with different types of artifacts to get a better estimate for the clean image.

Several intermediate reconstructions for a natural image in the test dataset are shown in Fig. 2.9. In fact, our approach generally does not introduce artifacts and errors like the other methods. Additionally, by considering the perception-distortion tradeoff, our approach also mitigates the side effects of smoothing that are prevalent in other methodologies, as discussed in [33]. This consideration allows us to strike a bal-

ance between preserving fine details in the reconstructed images while minimizing distortions, ultimately enhancing the perceptual quality of the results.

To evaluate the generalization capacity of different algorithms, the results for both natural and synthetic test images are presented separately in Table 2.1. The table shows that, although the DNNs were trained exclusively with natural images, the developed method achieves superior reconstruction performance for both natural and synthetic images, despite the distinct statistical properties of the latter.

Notably, the performance of the prDeep method declines significantly for synthetic images, which is anticipated since its reconstruction depends on a regularization prior learned from natural images. To highlight this, example reconstructions for a synthetic image from the test dataset are displayed in Fig. 2.10.

The table also reveals that the developed approach outperforms other methods across various noise levels ($\alpha = 2, 4$) in terms of reconstruction quality, despite being trained for a specific noise level ($\alpha = 3$). This indicates the robustness of the developed method to different noise conditions.

Phase retrieval algorithms are generally sensitive to initialization due to the inherent nonlinearity of the problem. To demonstrate the robustness of the developed approach to different initializations and image characteristics, PSNR and SSIM histograms are provided in Fig. 2.11 for the developed methods (with $\alpha = 3$). These histograms include reconstructions obtained from 236 distinct test images and 5 Monte Carlo runs, implying that 5 different initializations were used for each test image. The small spreads and high means clearly indicate the robustness of the developed approaches to varying initializations and image statistics.

The average runtime of each method is also listed in Table 2.1. Our methods not only outperform the other methods in terms of reconstruction quality but also exhibit computational efficiency comparable to the robust HIO initialization method, demonstrating both superior performance and efficiency.

Our findings also demonstrate that incorporating TTA during the reconstruction pipeline enhances performance even without additional training data. This suggests that TTA can take advantage of the inherent properties of the measurement model to improve

reconstruction accuracy, potentially reducing the reliance on extensive training datasets for specific noise distributions or image types. This observation warrants further investigation into TTA's role in generalizing deep learning methods for image reconstruction tasks.

It is important to mention that due to the realness and positiveness assumptions inherent in the measurement model, we inherently avoid the issue of trivial global phase shift ambiguity. While presenting the test results, we also addressed conjugate inversion ambiguity by comparing both the generated image and its flipped version with the ground truth image. This ensures that we capture the correct orientation of the reconstructed object.

However, the challenge of spatial circular shift ambiguity remains. Natural images generally possess a well-distributed intensity pattern across the known support, which helps to naturally break these shift symmetries. Interestingly, this issue has received limited discussion in previous phase retrieval literature, as evidenced by works such as [27], [81]. Notably, the initial HIO reconstruction is not susceptible to this specific ambiguity.

The compared methods disambiguate this circular shift ambiguity by using the ground-truth images, but, we did not deploy such a strategy.

But, our approach encounters limitations when dealing with certain unnatural test images, such as "E.Coli" and "Yeast." These images do not fully fit into the known support as observed in natural images. This can lead to multiple valid HIO reconstructions for the same observations, creating an ambiguity issue. While techniques like the shrinkwrap procedure [82] offer solutions to refine the support and address this ambiguity, we opted not to implement this step as our primary focus lies on natural image reconstruction.

Additionally, perceptual quality metrics, commonly employed to assess the fidelity of reconstructed images in human perception, are not presented in this work. While such metrics are valuable for evaluating reconstructions intended for human consumption, they often rely on deep learning models trained on natural color images. Since our focus is on grayscale phase retrieval and a suitable, widely-used perceptual quality

metric for this domain is not readily available, we primarily rely on established distortion metrics to quantify reconstruction performance.

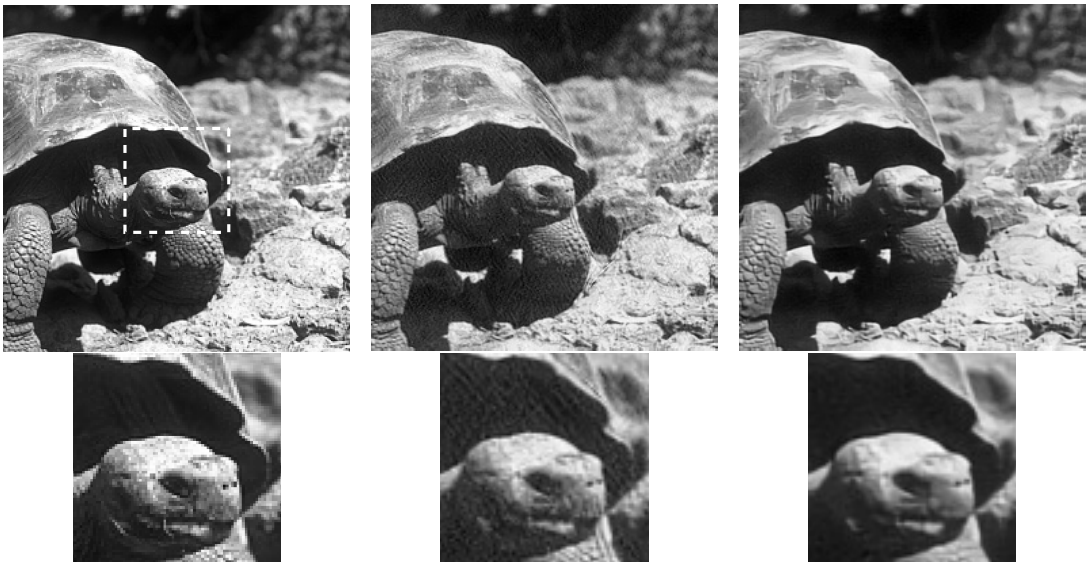
Table 2.1: Average reconstruction performances for 236 test images across 5 Monte Carlo runs.

$\alpha = 2$	Avg. PSNR (dB) ↑			Avg. SSIM ↑			Avg. runtime (sec.) ↓
(Avg. SNR: 33.24dB)	Overall	Natural	Unnatural	Overall	Natural	Unnatural	
HIO [19]	19.79	19.73	21.92	0.50	0.50	0.49	0.25
prDeep [10]	23.45	23.49	21.72	0.65	0.66	0.58	59.32
DIR [33]	23.61	23.60	24.02	0.72	0.72	0.73	21.59
MBwDDP [34]	24.87	24.86	25.56	0.74	0.74	0.74	24.11
Initialization stage-Small	20.64	20.55	24.25	0.53	0.53	0.57	0.45
prNet-Small	29.60	29.67	26.81	0.85	0.85	0.78	0.98
Main loop-Large	32.24	32.28	30.77	0.90	0.90	0.89	1.46
prNet-Large	32.38	32.42	30.77	0.90	0.90	0.89	1.50
prNet-Large (+TTA)	32.66	32.69	31.25	0.91	0.91	0.89	1.80
prNet-Large (+TTA D_4)	32.92	32.94	31.88	0.91	0.91	0.88	3.12
$\alpha = 3$	Avg. PSNR (dB) ↑			Avg. SSIM ↑			Avg. runtime (sec.) ↓
(Avg. SNR: 31.53dB)	Overall	Natural	Unnatural	Overall	Natural	Unnatural	
HIO [19]	18.92	18.89	20.34	0.43	0.43	0.43	0.27
prDeep [10]	22.06	22.09	20.91	0.59	0.59	0.54	59.41
DIR [33]	22.87	22.85	23.50	0.68	0.68	0.71	21.72
MBwDDP [34]	23.92	23.92	23.98	0.70	0.70	0.69	24.35
Initialization stage-Small	19.73	19.68	21.65	0.46	0.46	0.45	0.47
prNet-Small	28.08	28.13	25.93	0.80	0.81	0.70	1.03
Main loop-Large	30.17	30.24	27.79	0.86	0.86	0.78	1.48
prNet-Large	30.22	30.28	27.79	0.86	0.87	0.79	1.52
prNet-Large (+TTA)	30.52	30.57	27.88	0.87	0.87	0.79	1.82
prNet-Large (+TTA D_4)	30.73	30.81	27.76	0.87	0.87	0.76	3.13
$\alpha = 4$	Avg. PSNR (dB) ↑			Avg. SSIM ↑			Avg. runtime (sec.) ↓
(Avg. SNR: 30.24dB)	Overall	Natural	Unnatural	Overall	Natural	Unnatural	
HIO [19]	18.52	18.48	19.80	0.39	0.39	0.40	0.28
prDeep [10]	20.69	20.70	20.38	0.53	0.53	0.51	59.68
DIR [33]	21.80	21.77	22.79	0.62	0.62	0.69	21.95
MBwDDP [34]	22.41	22.39	23.09	0.63	0.63	0.65	24.43
Initialization stage-Small	19.12	19.08	20.63	0.41	0.41	0.41	0.48
prNet-Small	26.69	26.75	24.48	0.76	0.76	0.67	1.02
Main loop-Large	28.29	28.35	25.98	0.81	0.81	0.72	1.48
prNet-Large	28.29	28.35	25.91	0.81	0.81	0.72	1.52
prNet-Large (+TTA)	28.56	28.61	26.49	0.82	0.82	0.73	1.81
prNet-Large (+TTA D_4)	28.74	28.80	26.50	0.83	0.83	0.73	3.12

(a) Ground truth

(b) prDeep, PSNR:25.35, SSIM:0.71

(c) DIR, PSNR:26.49, SSIM:0.73



(d) MBwDDP, PSNR:27.87, SSIM:0.77

(e) prNet-Small, PSNR:28.67, SSIM:0.89

(f) prNet-Large (+TTA), PSNR:30.88, SSIM:0.92

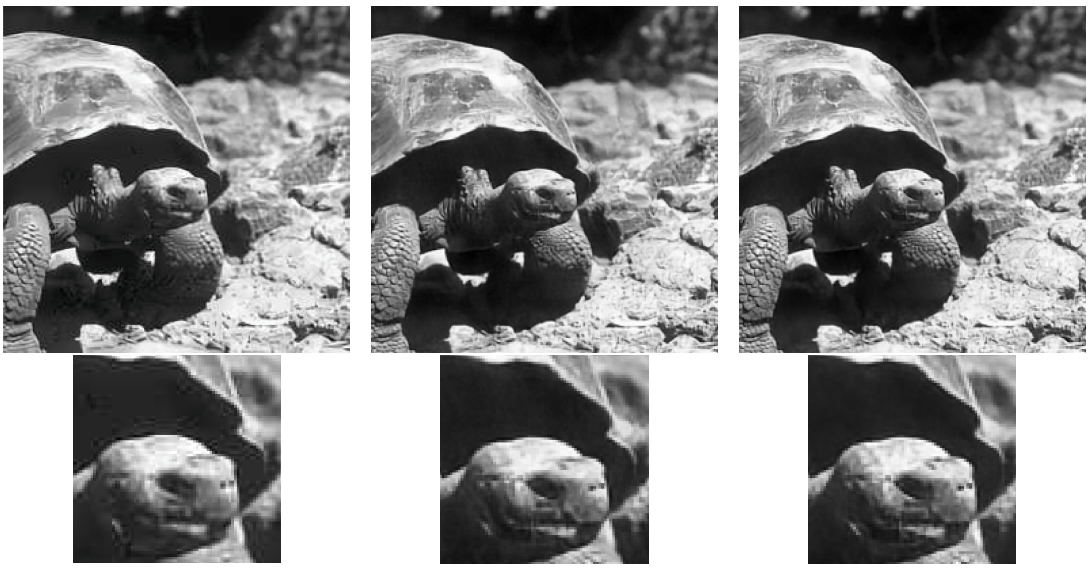


Figure 2.7: The outputs of various algorithms for the "Turtle" test image subjected to $\alpha = 3$ noise (SNR=31.89dB).

(a) Ground truth

(b) prDeep, PSNR:23.44, SSIM:0.66

(c) DIR, PSNR:20.37, SSIM:0.54

(d) prNet-Small, PSNR:27.12, SSIM:0.88

(e) prNet-Large, PSNR:30.87, SSIM:0.93

(f) prNet-Large (+TTA), PSNR:31.23, SSIM:0.93

Figure 2.8: The outputs of various algorithms for the "Cameraman" test image subjected to $\alpha = 3$ noise (SNR=31.61dB).

(a) Ground truth

(b) Initialization stage-Small, PSNR:19.19, SSIM:0.45

(c) prNet-Small, PSNR:33.51, SSIM:0.94

(d) Main loop-Large, PSNR:35.19, SSIM:0.95

(e) prNet-Large, PSNR:35.27, SSIM:0.95

(f) prNet-Large (+TTA), PSNR:35.50, SSIM:0.95

Figure 2.9: Intermediate reconstruction results from the developed approaches for the "Woman" test image at a noise level of $\alpha = 3$ (SNR=32.09dB).

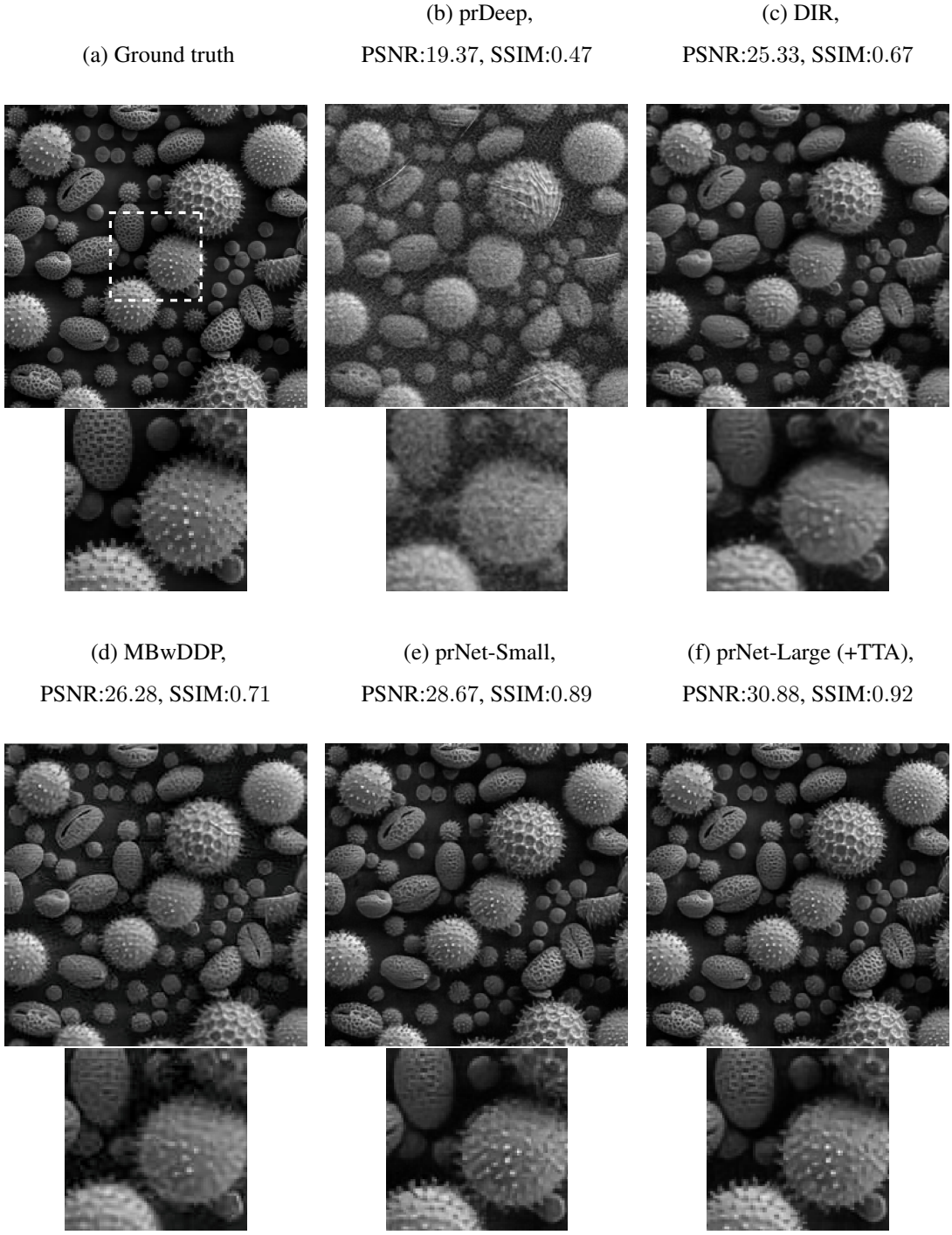


Figure 2.10: The outputs of various algorithms for the out-of-domain "Pollen" test image subjected to $\alpha = 3$ noise (SNR=28.10dB).

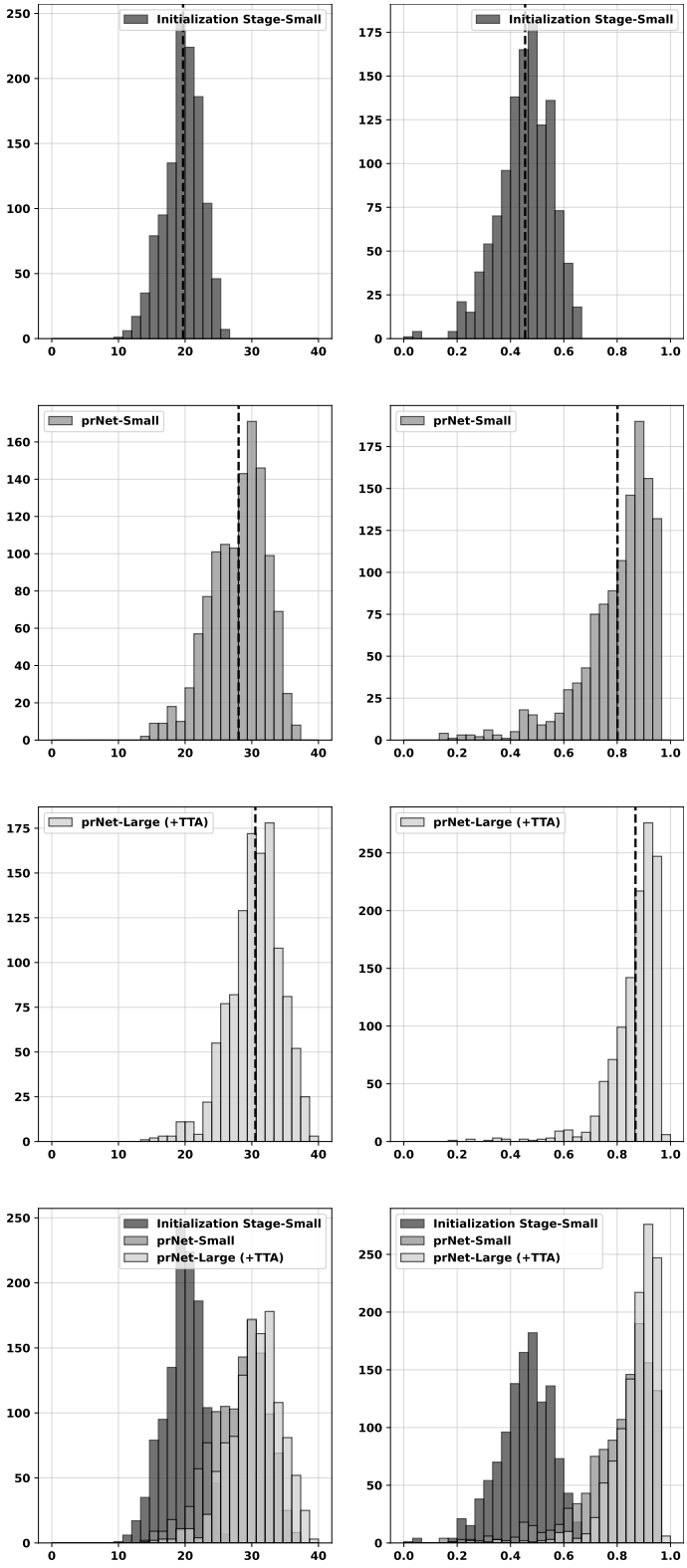


Figure 2.11: The histograms of PSNR (left column) and SSIM (right column) for the reconstructions produced by various methods across 236 test images and 5 Monte Carlo runs for the $\alpha = 3$ scenario. Vertical dashed lines indicate the mean PSNR and SSIM values. Overlapping histograms for each column are shown at the bottom.

2.5 Conclusion

This chapter presents a novel approach to phase retrieval based on Langevin dynamics for posterior sampling. Due to Tweedie's formula, it can also be seen as an application of diffusion models to the phase retrieval problem [54]. Unlike existing methods trained solely with distortion metrics, which often suffer from overly smooth outputs, our approach considers the perceptual quality-distortion tradeoff, resulting in reconstructions with high fidelity and low distortion metrics.

We introduce two deep learning architectures based on this sampling procedure: prNet-Small and prNet-Large. prNet-Small offers a compact and efficient solution. prNet-Large, an extension utilizing multiple outputs, enhances robustness to initialization by incorporating diverse starting points, ultimately facilitating artifact removal.

Both methods start with initial HIO estimates and iteratively refine them through a denoising-data consistency-noise addition cycle. Rather than assuming a diffusion process and training based on this assumption, the training process exactly utilizes the output of the previous iterations, similar to unrolled optimization algorithms, minimizing the loss between the reconstruction at a given timestep and the corresponding ground truth image. During training, random timesteps are incorporated progressively. This is achieved by placing greater emphasis on the later stages of the reconstruction process as the training epochs progress.

prNet-Large model uses an extra denoiser to combine the multiple outputs of its main loop. The loss function for this extra denoiser includes an extra Wasserstein loss term together with the reconstruction loss to also compensate for the perceptual quality of the generated images.

Also, by using the properties of the phase retrieval problem, test time augmentation is applied to the overall prNet-Large pipeline to improve the reconstruction performance further with a little extra time and no training requirement.

Our developed methods were rigorously tested through various numerical simulations and benchmarked against both classical and state-of-the-art techniques. The findings

reveal that our approach is highly effective, adding only minimal computational overhead compared to the robust initialization method based on HIO. It achieves superior reconstruction performance and demonstrates greater resilience to variations in initialization, image characteristics, and noise levels.

To conclude, the methods we have developed offer cutting-edge reconstruction capabilities and computational efficiency for tackling the phase retrieval problem. We believe that initiating diffusion processes with a preliminary estimate and integrating denoisers with model-based approaches, as shown in this chapter, could be pivotal in advancing more reliable algorithms for both phase retrieval and broader nonlinear inverse problems.

CHAPTER 3

INDI-PR: ENHANCING FOURIER PHASE RETRIEVAL THROUGH INVERSION BY DIRECT ITERATION

3.1 Introduction

In recent years, deep learning has emerged as a powerful tool for solving various inverse problems in imaging, including phase retrieval. These data-driven approaches, particularly those utilizing deep neural networks (DNNs), have shown remarkable success in directly reconstructing images from measurements or refining initial estimates from classical methods. However, existing deep learning solutions for PR often grapple with limitations such as domain shift, lack of interpretability, or the necessity for extensive parameter tuning. Moreover, most deep learning methods, including those based on the unrolling of iterative algorithms, face significant challenges, such as lengthy training processes and inefficient use of computational resources [24].

Unfolding methods, which are designed to mimic traditional iterative reconstruction algorithms through a series of trainable network layers, often suffer from prolonged training times and significant computational overhead. Furthermore, many current methods commence the recovery from a state of random noise, which can lead to inefficient utilization of denoiser capacity and extended convergence times, ultimately limiting their practical applicability.

To address these gaps, our work introduces a novel approach within the Inversion by Direct Denoising (InDI) [52] framework to enhance the classical Fourier PR. This methodology marks a significant departure from traditional methods by leveraging advanced denoising strategies combined with novel initialization and ensembling techniques. By initiating the recovery process from a plausible estimate rather than

random noise, our approach more efficiently utilizes the denoiser's model capacity and significantly reduces training time compared to conventional unrolling methods.

The primary contributions of our research include:

- We have integrated a novel accelerated error reduction algorithm into our initialization strategy, significantly enhancing the robustness and speed of convergence for the phase retrieval process.

- Diverging from other diffusion-based methods, our image-to-image framework starts with a plausible initial estimate and refines it, optimizing the denoiser's capacity and reducing overall training duration.

- Our ensembling technique combines multiple reconstructions to improve distortion metrics, but also substantially enhancing the perceptual quality of the reconstructed images.

The techniques developed in this study not only advance the field of classical Fourier phase retrieval but also open promising avenues for their application to other types of phase retrieval challenges. Our methods demonstrate superior performance compared to both classical and contemporary techniques, underscoring their efficacy in addressing the intrinsic challenges of PR. Furthermore, the hybridization of denoisers with model-based approaches holds promise for developing robust and reliable stochastic nonlinear inverse problem solvers with broad applications beyond phase retrieval.

The structure of this chapter is organized as follows: Section 3.2 reviews existing research that informed the development of our approach. The methodology of our developed approach, including the novel aspects of the Inversion by Direct Denoising (InDI) framework, is detailed in Section 3.3. Comparative analyses of our method against both classical techniques and contemporary advancements are presented in Section 3.4. Finally, Section 3.5 consolidates our key findings and articulates prospective avenues for future research in this dynamic field.

3.2 Related Works

3.2.1 The Geometric Interpretation of Classical Iterative Methods for Phase Retrieval

The geometric interpretation of the Error Reduction (ER) and Hybrid Input-Output (HIO) algorithms provides a clear visualization of their operational mechanics in the context of phase retrieval. Both algorithms utilize iterative projections between constraint sets defined in different domains, but they differ significantly in their approach to managing deviations from these constraints.

The ER algorithm applies consecutive projection operations to refine the estimate iteratively, aligning it within the intersection of spatial and Fourier magnitude constraints. Mathematically, this is represented by:

$$\mathbf{x}_{k+1} = \mathcal{P}_s \mathcal{P}_f \mathbf{x}_k \tag{3.1}$$

where $\mathcal{P}_f$ and $\mathcal{P}_s$ are projection operators enforcing Fourier magnitude and spatial domain constraints, respectively. Geometrically, this sequence of projections directs the estimate towards the intersection of the constraint sets in a straightforward, stepwise manner.

Conversely, HIO incorporates a reflective step to handle violations of spatial constraints, allowing for correction of the trajectory during iterations. The iteration formula for HIO is expressed as:

$$\mathbf{x}_{k+1} = \left[\frac{\mathbf{I} + \mathcal{R}_s \mathcal{R}_f}{2} + (1-\beta)(\mathbf{I} - \mathcal{P}_s)\mathcal{P}_f \right] \mathbf{x}_k \tag{3.2}$$

Here, $\mathcal{R}_s$ and $\mathcal{R}_f$ denote reflection operators related to the spatial and Fourier constraints, respectively. This formulation introduces a dynamic adjustment to the trajectory, allowing the method to navigate around potential local minima and avoid stagnation, a common limitation in simpler projection methods.

The ER algorithm guarantees a direct approach toward the solution by closely following the constraints, while HIO allows for a more explorative strategy, potentially circumventing issues like local minima through its reflective and corrective steps.

These interpretations underscore the distinct pathways each algorithm takes in the constrained solution landscape of phase retrieval.

Notably, when $\beta = 1$, HIO becomes equivalent to the Douglas-Rachford algorithm. This equivalence is particularly significant as the Douglas-Rachford algorithm is known for its efficacy in handling nonconvex feasibility problems, such as phase retrieval. However, while HIO offers advantages in avoiding local minima due to its more explorative update strategy, it can exhibit challenges such as spiraling dynamics [83].

3.2.2 Image-to-Image Pipelines for Inverse Problems

Unrolled methods for solving inverse problems are noted for their computational and memory inefficiencies as well as slow training, primarily due to their extensive computational requirements and the iterative refinement they employ [84]. In contrast, classical diffusion pipelines for addressing inverse problems typically initiate the restoration from a random noise image. This starting point is less than ideal because it does not begin with a crude reconstruction that can be generated by classical algorithms, which could potentially facilitate faster convergence and avoid the waste of the denoiser's model capacity.

On the other hand, image-to-image pipelines initiate the process with a warm start procedure. By starting with a plausible estimate rather than a complete noise image, they facilitate training since it should only learn to refine this initial estimate. This "image-to-image" instead of "noise-to-image" idea is prevalent in the literature [52], [69]–[71].

One such image-to-image pipeline, the Inversion by Direct Denoising (InDI) method, starts from a basic estimate of the image. This initial approach not only conserves denoiser capacity but also enhances the training process by defining a specific diffusion process that simplifies model training [52].

The stochastic version of the Inversion by Direct Denoising (InDI) method employs a sophisticated approach to image restoration by integrating denoising diffusion probabilistic models into its framework. This methodology capitalizes on the incremental improvement of image quality through iterative denoising, each modified by stochas-

tic perturbations, which are essential for handling various degradation levels and ensuring robustness in the denoising process.

The training of the denoising model within the InDI framework is strategically designed to cope with varied noise levels, introduced into the training data through a simulation of the noise degradation process. The model employs the equation for the intermediate degraded image:

$$\mathbf{x}_t = (1-t)\mathbf{x} + t\mathbf{z} + t\sigma_t\boldsymbol{\epsilon}, \tag{3.3}$$

where $\mathbf{x}$ represents the clean target image, $\mathbf{z}$ is the low quality input, t ranges from 0 to 1, indicating the transition from clean to noisy image, σ_t varies with t as the standard deviation of noise, and $\boldsymbol{\epsilon}$, following a standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, introduces stochasticity. This formulation prepares the model to reverse noise effects by optimizing the neural network parameters $\boldsymbol{\theta}$ to minimize:

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p(\mathbf{x},\mathbf{y})} \left[\mathbb{E}_{t\sim p(t)} \left[\|\text{Denoiser}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) - \mathbf{x}\|^2 \right] \right], \tag{3.4}$$

which is the mean squared error between the denoised image and the original clean image across randomly sampled noise levels t.

Once trained, the model uses the denoising function to iteratively restore the noisy image towards its original state. The guiding recurrence relation for this inference is:

$$\widehat{\mathbf{x}}_{t-\tau} = \left(1 - \frac{\tau}{t}\right) \widehat{\mathbf{x}}_t + \frac{\tau}{t} \,\text{Denoiser}(\widehat{\mathbf{x}}_t, t) + (t-\tau)\sqrt{\sigma_{t-\tau}^2 - \sigma_t^2}\boldsymbol{\epsilon}, \tag{3.5}$$

where τ is a small step back in time from t, enhancing the restoration precision at each step. The term $\sigma_{t-\tau}^2 - \sigma_t^2$ reflects the decrease in noise variance, aiding in the gradual restoration process.

The function $\text{Denoiser}(\widehat{\mathbf{x}}_t, t)$ calculates the expected clean image given the current noisy estimate, mirroring the conditional expectation of the posterior distribution $\mathbb{E}[\mathbf{x}_{t-1} \mid \mathbf{x}_t]$. This function, optimized during training, encapsulates the core denoising capability of the model, aimed at minimizing the reconstruction error.

The InDI method introduces an innovative approach to supervised image restoration that mitigates the "regression to the mean" effect, yielding images that are more realistic and detailed compared to traditional regression-based techniques. This method

improves image quality incrementally in small steps, similar to generative denoising diffusion models. Traditional single-step regression models often produce averaged outputs that lack detail and realism due to the ill-posed nature of the problem, where multiple high-quality images can plausibly reconstruct a given low-quality input. InDI's strength lies in its iterative refinement process, which enhances perceptual quality by gradually improving the image. Unlike generative denoising diffusion models that need prior knowledge of the degradation process, InDI learns the restoration directly from paired examples of low-quality and high-quality images. This approach is applicable to various image degradation scenarios, making it a versatile and powerful solution for image restoration tasks [52].

The InDI framework, under certain conditions, is equivalent to other image-to-image pipelines such as Schrödinger Bridge [85], [86] and Cold Diffusion [52], [70]. This equivalence underscores the versatile and robust nature of the InDI method, aligning its operational dynamics with the established methodologies that worked well for other image reconstruction problems [87].

3.3 Developed Method

3.3.1 Initialization Procedure

The challenge of phase retrieval is exacerbated by the inherent nonlinearity and non-convexity of the problem. These characteristics render the algorithm highly sensitive to how the reconstruction process is initialized. In our method, to mitigate these challenges and improve the robustness of the reconstruction, we adopt a sophisticated initialization strategy that builds upon the principles described in [10]. This strategy involves a hybrid approach combining the Hybrid Input-Output (HIO) method with Error Reduction (ER) techniques, further enhanced by an acceleration mechanism.

Initially, the HIO method is employed using multiple random initializations. This step is crucial as it explores various potential starting points in the solution space, each initialized with a different random phase. The HIO algorithm is run for a small predefined number of iterations (denoted as k), which allows each initialization to

evolve without significant computational overhead. This initial exploration aims to identify promising regions in the search space and is highly parallelizable.

Following the initial HIO iterations, we evaluate each result by computing the residual $\|\mathbf{y} - |\tilde{\mathbf{F}}\mathbf{x}|\|^2$. The k initializations yielding the lowest residuals are selected for further refinement. These chosen estimates undergo additional HIO+ER processing, this time for an extended number of iterations n, to enhance the fidelity of the reconstructions.

To further refine the output and accelerate convergence towards a high-fidelity reconstruction, the selected initialization then undergoes a combined ER and HIO regimen. This regimen is structured in cycles: for a set number of iterations, the reconstruction alternates between applying the HIO constraint and the ER constraint. Notably, during the ER phase, an acceleration step is incorporated every few iterations, as given in Fig. 3.1 and Algorithm 2. This acceleration involves a dynamic adjustment of the current estimate by leveraging a convex combination of the current and previous estimates, moderated by a scaling factor ζ. This step is critical as it introduces a momentum-like effect, propelling the reconstruction towards the ground truth more effectively by reducing stagnation in local minima.

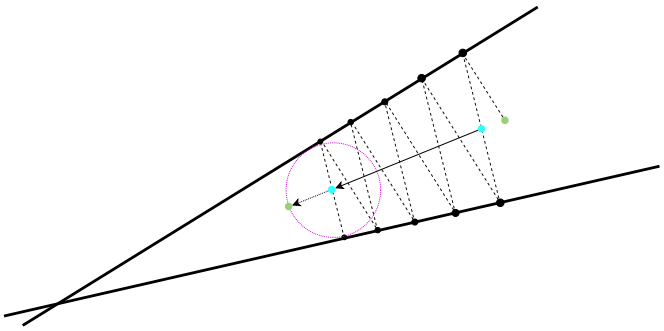


Figure 3.1: Geometric interpretation of the acceleration mechanism used during the ER phase in the initialization procedure.

Empirically, this initialization and iterative refinement approach has demonstrated superior performance, particularly in terms of achieving lower residuals and higher image quality. The inclusion of an acceleration mechanism during the ER steps further enhances the efficiency of the reconstruction process, allowing for faster convergence while maintaining the integrity of the reconstructed image as they reliably reduce the

error in each iteration.

In summary, our methodological framework for phase retrieval initialization not only adheres to established algorithms but also innovates by integrating a tailored acceleration technique during the ER phases. This strategy significantly improves the robustness and effectiveness of the phase retrieval process, making it well-suited for complex imaging scenarios where traditional methods struggle.

Algorithm 2 Proposed accelerated ER (AER) algorithm

for $n = 1$ to K **do**
  $\mathbf{x}'_n \leftarrow \mathcal{P}_F \mathbf{x}_n$
  $\mathbf{x}_{n+1} \leftarrow \mathcal{P}_S \mathbf{x}'_n$

  if $n = -1 \pmod t$ **then**
    $\mathbf{c}_n \leftarrow \frac{1}{2}(\mathbf{x}'_n + \mathbf{x}_{n+1})$
    $\mathbf{a} \leftarrow \frac{\mathbf{c}_n - \mathbf{c}_{n-t}}{\|\mathbf{c}_n - \mathbf{c}_{n-t}\|}$
    $r \leftarrow \frac{1}{2}\|\mathbf{x}'_n - \mathbf{x}_{n+1}\|$
    $\mathbf{x}_{n+1} \leftarrow \mathbf{c}_n + \zeta r \mathbf{a}$

3.3.2 Iterative Refinement through Inversion by Direct Iteration

As mentioned before, our initialization procedure produces k different outputs for the same measurement $\mathbf{y}$. These are crude estimates of the unknown image and are denoted by $\{\hat{\mathbf{x}}_{\text{init}}^{(m)}\}_{m=1}^{k}$. In the context of the InDI framework, we start the iterative refinement with the mean of these multiple initial reconstructions, mathematically, $\mathbf{z} = \frac{1}{k}\sum_m \hat{\mathbf{x}}_{\text{init}}^{(m)}$.

To counteract the information loss typically associated with this averaging, our empirical research has demonstrated that conditioning the denoiser on multiple initial reconstructions significantly improves reconstruction performance. Specifically, rather than using the standard input in the original InDI framework, which is $\mathrm{Denoiser}(\widehat{\mathbf{x}}_t, t)$, we now incorporate a set of k initial reconstructions. Consequently, the denoiser is now conditioned to operate as $\mathrm{Denoiser}(\widehat{\mathbf{x}}_t, t, \{\hat{\mathbf{x}}_{\text{init}}^{(m)}\}_{m=1}^{k})$ at each step of the iterative refinement. This configuration means that the denoiser takes $k+1$ inputs – k from

the multiple initial reconstructions and one from the classical InDI current estimate $\widehat{\mathbf{x}}_t$ and produces one estimate. This approach utilizes additional context from the initial estimates, significantly enhancing the accuracy and efficacy of the image restoration process.

The proposed pipeline outlined in Algorithm 3 and Fig. 3.2 employs an efficient utilization of the denoiser's model capacity. In each iteration, after denoising and data consistency with HIO, we also add a Gaussian noise following the InDI method. Only iteratively denoising the estimate without physics-informed blocks can produce outputs that are not compatible with the measurements. However, despite HIO's advantages, noise and local minima can still introduce artifacts in reconstructions. To address this, we employ an iterative denoising-data consistency approach also used in many unrolling methods [37]. This scheme aims to escape local minima and reduce artifacts, leading to improved results.

For training our pipeline, we follow the InDI training strategy based on a carefully defined noising process. An example of the gradual noising used during training can be seen in Fig. 3.3.

As the denoising component of our pipeline for the phase retrieval problem, we employed a customized UNet architecture, as depicted in Fig. 3.4. This well-established framework is renowned for its efficacy in image restoration tasks. Our implementation of UNet uniquely incorporates timestep information as an additional input, which is intricately linked to the noise level of the input image. The denoiser operates by estimating the residual, thereby facilitating the refinement of the reconstructed image by focusing solely on the discrepancy between the noisy input and the desired clean output.

Our customized UNet architecture includes several enhancements, most notably the integration of attention mechanisms within the convolutional blocks. These attention mechanisms enable the network to selectively focus on the most relevant parts of the input image, thereby enhancing its capacity to capture intricate details and effectively suppress noise. This selective focus is particularly beneficial in scenarios where noise levels vary across different regions of the image, a common challenge in phase retrieval tasks. Additionally, the use of attention mechanisms helps preserve

high-frequency details that are crucial for accurate phase reconstruction.

Algorithm 3 Overall pipeline: InDI-PR

Input: $\mathbf{y}$, $T, K, \sigma_i, \beta, \lambda \in \mathbb{R}^T$ is learnable (initially, a logarithmically increasing vector)

Output: $\mathbf{z}_T^{(0)}$

Initialization:

$\{\hat{\mathbf{x}}_{\text{init}}^{(m)}\}_{m=1}^k \leftarrow$ Initialization procedure($\mathbf{y}$)

$\mathbf{w} \leftarrow \mathcal{N}(\mathbf{0}, \mathbf{I})$

$\mathbf{x}'_{T+1} \leftarrow \frac{1}{k}\sum_m \hat{\mathbf{x}}_{\text{init}}^{(m)} + \sigma_T \mathbf{w}$

Main loop:

for $i = T$ to 1 **do**

  $\mathbf{x}_i \leftarrow$Denoiser($\mathbf{x}'_{i+1}$, i, $\{\hat{\mathbf{x}}_{\text{init}}^{(m)}\}_{m=1}^k$)

  $\mathbf{z}_i^{(0)} \leftarrow \mathbf{O_{mn}}\mathbf{x}_i$

  $\mathbf{y}_i{}' \leftarrow \lambda_i \mathbf{y} + (1-\lambda_i)|\mathbf{F}\mathbf{z}_i^{(0)}|$

  for $k = 1$ to K **do**

    $\mathbf{z}_i^{(k)'} \leftarrow \Re\{\mathbf{F^{-1}}[\mathbf{y}_i{}' \odot \frac{\mathbf{F}\mathbf{z}_i^{(k-1)}}{|\mathbf{F}\mathbf{z}_i^{(k-1)}|}]\}$

    $\gamma \leftarrow$ the set of indices where $\mathbf{z}_i^{(k)'}$ violates space domain constraints (e.g., support and non-negativity)

    $$\mathbf{z}_i^{(k)}[n] \leftarrow \begin{cases} \mathbf{z}_i^{(k)'}[n] & , n \notin \gamma \\ \mathbf{z}_i^{(k-1)}[n] - \beta \mathbf{z}_i^{(k)'}[n] & , n \in \gamma \end{cases}$$

  $\epsilon \leftarrow \mathcal{N}(\mathbf{0}, \mathbf{I})$

  $\mathbf{x}'_i \leftarrow \frac{1}{i}\sqrt{\frac{n}{m}}\mathbf{O^T_{mn}}\mathbf{z}_i^{(K)} + \left(1 - \frac{1}{i}\right)\mathbf{x}'_{i+1} + \frac{i-1}{T}\sqrt{\sigma_{i-1}^2 - \sigma_i^2}\epsilon$

return $\mathbf{x}'_1$

3.3.3 Ensembling Scheme

We can process the initialization outputs $\{\hat{\mathbf{x}}_{\text{init}}^{(m)}\}_{m=1}^k$ with an equivariant transform $\mathcal{T}$, such as flipping, to easily produce different set of initializations $\{\mathcal{T}(\hat{\mathbf{x}}_{\text{init}}^{(m)})\}_{m=1}^k$. Then, we can iteratively refine these two different initialization outputs and get two different final outputs. We can ensemble these two final results by simple averaging,

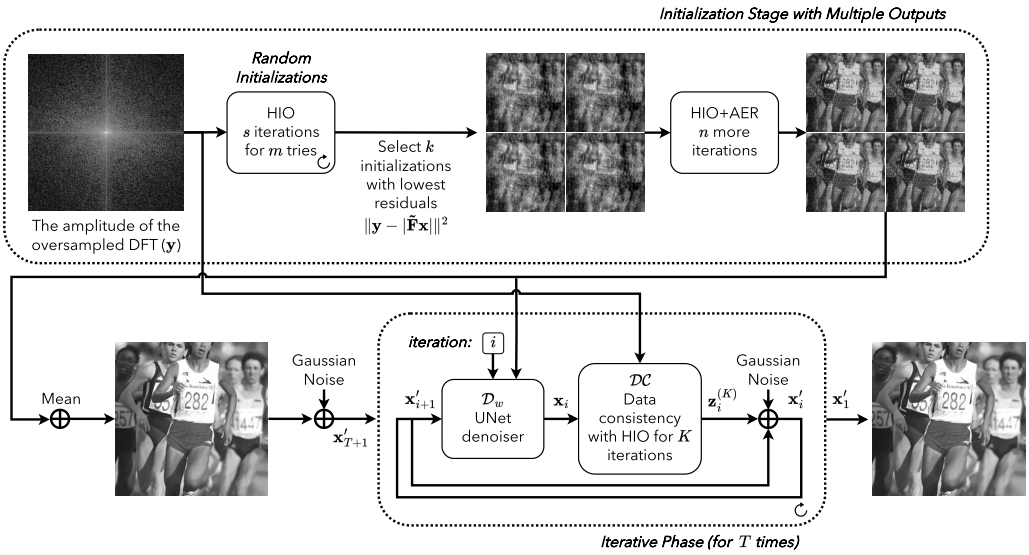


Figure 3.2: The overall pipeline of InDI-PR.

Figure 3.3: An example of the defined gradual process used during training. The timestep is increasing from left to right. The rightmost image ($t = 1$) corresponds to the output of the initialization procedure, and the leftmost image ($t = 0$) corresponds to the clean image.

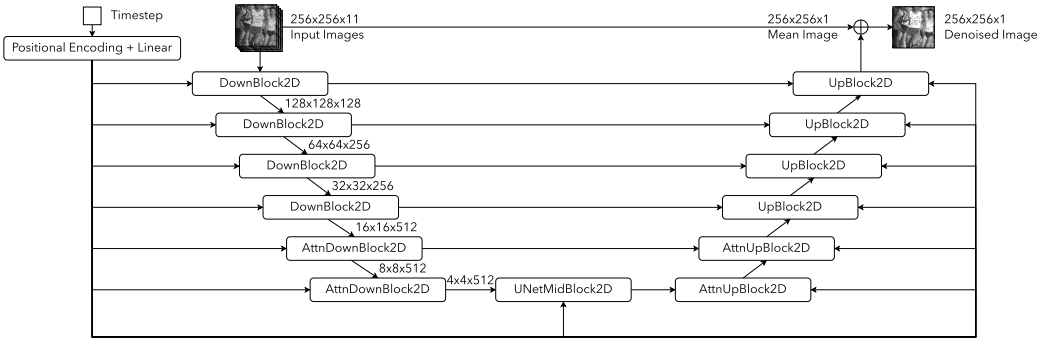


Figure 3.4: Architecture of the UNet denoiser with multiple input images and a timestep input producing one denoised image.

$\hat{\mathbf{x}}_{\text{final}}^{(\text{combined})} = \frac{1}{2}\left(\hat{\mathbf{x}}_{\text{final}}^{(\text{original})} + \mathcal{T}^{-1}(\hat{\mathbf{x}}_{\text{final}}^{(\text{transformed})})\right)$. This augmentation process is depicted in Fig. 3.5.

Let produce p different such combined results, $\hat{\mathbf{x}}_{\text{final}}^{(\text{combined})}$, by starting our algorithm from scratch. As our algorithm has stochastic components, such as the random initial phase for the initialization procedure and Gaussian noise in the iterative refinement stage, each output will be different from the other ones. Then, we can again combine these p different final results by simple averaging.

As the result of such an ensembling scheme, effectively, we are combining $2p$ samples $\{\hat{\mathbf{x}}_{\text{final}}^{(\text{q})}\}_{q=1}^{2p}$ from the posterior distribution $p(\mathbf{x}|\mathbf{y})$. As p grows, the ensemble average, $\bar{\mathbf{x}}_{\text{final}} = \frac{1}{2p}\sum_q \hat{\mathbf{x}}_{\text{final}}^{(\text{q})}$, converges to the MMSE estimate, and we expect to see better distortion metrics.

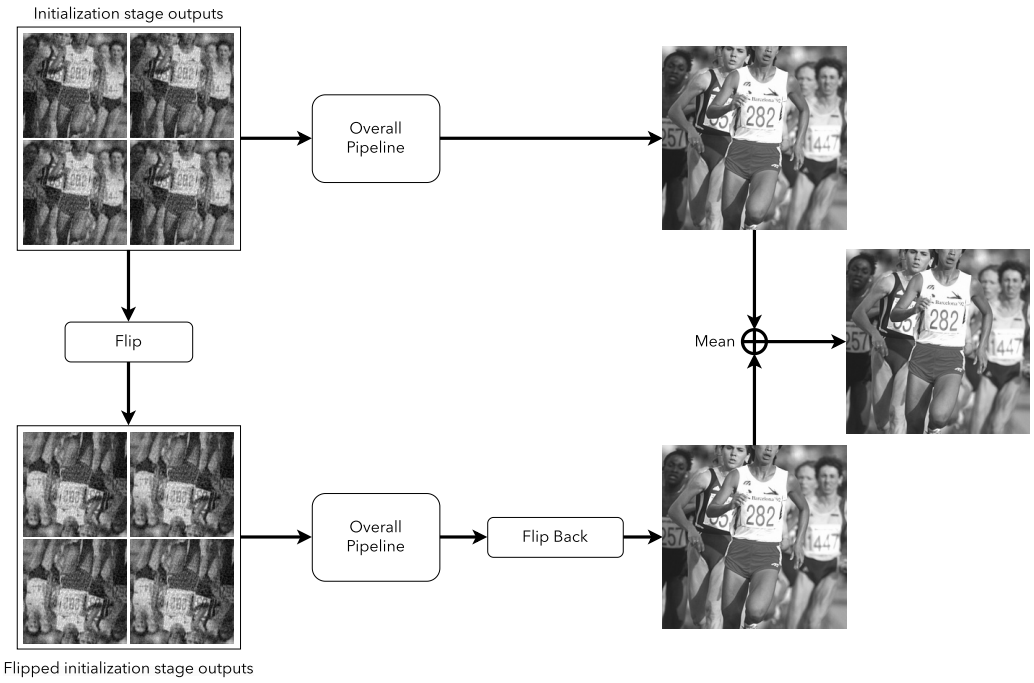


Figure 3.5: Test time augmentation with an equivariant transform.

3.4 Results

To evaluate the performance of our method, we conducted numerical simulations using a large image dataset. We compare the reconstruction performance against both classical and state-of-the-art phase retrieval methods.

To assess the algorithms' robustness to noise, generalization capacity, and computational cost, we investigate their reconstruction performance in two distinct image categories: natural and unnatural. This categorization allows us to evaluate the methods' ability to handle real-world scenarios (natural images) and potentially more challenging, synthetic long-tail samples (unnatural images).

For the training phase, exclusively natural images are utilized. The training dataset comprises $44,000$ natural images, including 200 training and 100 validation images from the Berkeley segmentation dataset (BSD) [72], $41,400$ images selected from the ImageNet database [73], [74], and $2,300$ images randomly chosen from the Waterloo Exploration Database [75].

During the testing phase, both natural and synthetic images are employed. This test dataset, previously used in [33], [34], contains 236 images, which include 230 natural and 6 synthetic images. Specifically, the dataset consists of 200 test images from BSD, 24 images from the Kodak dataset [76], and 6 natural and 6 synthetic images sourced from [10]. The synthetic image subset features images obtained from scanning electron microscopes and telescopes. All images have pixel values ranging from 0 to 255 and are of size 256×256.

The noisy Fourier measurements were generated using Eq. 1.2, with the average SNR values presented in Table 3.1 (where $\text{SNR} = 10\log(\left\||\tilde{\mathbf{F}}\mathbf{x}|^2\right\|_2 / \left\|\mathbf{y^2} - |\tilde{\mathbf{F}}\mathbf{x}|^{\mathbf{2}}\right\|_2)$).

Despite the training being conducted solely with natural images, the developed pipeline was evaluated on both natural and synthetic images to assess its generalization capabilities.

Decoupled weight decline regularization [78] is used for the training MSE loss optimization together with cosine annealing with linear warming [79]. The developed method is implemented by PyTorch and tested on a NVIDIA A100 80GB PCIe GPU.

The total training time was about 60 hours (27 epochs), respectively.

In the initialization stage, the HIO method was first run with $m = 100$ different random initializations for $s = 50$ iterations. Then, reconstructions with the lowest residuals was used for AER+HIO run for $n = 1700$ iterations with $\zeta \approx 1$. Thus, as the output of this initialization procedure, $k = 10$ multiple outputs are generated from the best $k = 10$ initializations with the lowest residuals among the $m = 100$ different random initializations.

3.4.1 Comparison with Other Methods

The evaluation of our newly developed approach involved comparing its reconstruction accuracy against original images using the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM) [80]. For a comprehensive comparison, results from the same test set were also generated using existing methods, namely prDeep [10], the classical HIO algorithm [18], DIR [33], and MBwDDP [34].

Table 3.1 displays the average reconstruction performance for 236 test images subjected to 5 Monte Carlo runs, with varying levels of Poisson noise ($\alpha = 2, 3, 4$). The table illustrates that our methods not only outperform the comparison group in both PSNR and SSIM metrics but also maintain computational efficiency comparable to robust initialization procedures based on HIO. The incremental performance enhancements at different stages of our algorithms are documented, emphasizing the efficacy of our approach.

Further analysis reveals that our approach maintains superior reconstruction quality across various noise levels ($\alpha = 2, 4$), indicating robustness even though the algorithms were initially optimized for a noise level of $\alpha = 3$. This adaptability is a testament to the resilience and versatility of our methods.

Visually, the superiority of our approach is demonstrated in Figs. 3.6 and 3.7, with additional examples available in Appendix B. The InDI-PR pipeline notably excels in removing HIO artifacts and preserving image characteristics, which is critical for maintaining the integrity of the reconstructed images.

By incorporating considerations of the perception-distortion tradeoff, our approach effectively mitigates the common smoothing artifacts prevalent in other methods. This strategy ensures a delicate balance between minimizing distortions and preserving fine details, thus enhancing the perceptual quality of the images, as discussed in [33].

Table 3.1 presents the performance results for different algorithms across natural and unnatural test images, allowing for an assessment of each method's generality. Notably, even though our deep neural networks (DNNs) were trained exclusively on natural images, our method achieves the best reconstruction performance for both image types. This demonstrates an impressive capability to generalize beyond the training data to images with differing statistical properties.

Despite the overall success, our method exhibits occasional shortcomings in the structural similarity index (SSIM) for unnatural images, although it consistently excels in terms of peak signal-to-noise ratio (PSNR). These discrepancies highlight potential areas for improvement, particularly in how our method handles the specific textural elements of unnatural images.

The prDeep method, in contrast, shows a significant drop in performance when processing unnatural images. This decline is anticipated, as its reconstruction relies heavily on a regularization prior tailored to the characteristics of natural images. For a visual comparison, sample reconstructions of an unnatural image from the test dataset are depicted in Fig. 3.8.

It is crucial to address the fact that our measurement model's assumptions of realness and positiveness inherently mitigate the trivial global phase shift ambiguity. Moreover, in our analysis, we confront the conjugate inversion ambiguity by comparing both the original and flipped versions of the generated image with the ground truth, ensuring accurate orientation alignment of the reconstructed objects.

However, challenges persist with spatial circular shift ambiguity, particularly notable in unnatural images such as "E.Coli" and "Yeast," which do not conform well to the known support pattern typical of natural images. This misalignment can lead to multiple valid reconstructions using the HIO algorithm, introducing notable ambiguities.

Previous literature on phase retrieval has only sparingly discussed this issue, with few exceptions like the studies by [27], [81]. While methods like the shrinkwrap procedure [82] are known to refine support and reduce ambiguities, we chose not to implement this step, focusing instead on the reconstruction of natural images.

Our decision not to use ground truth images to disambiguate circular shift, unlike the other compared methods, partly explains the lower performance observed in certain cases. This strategic choice highlights a tradeoff between methodological simplicity and the potential for increased error in specific contexts.

Table 3.1: Average reconstruction performances for 236 test images across 5 Monte Carlo runs.

$\alpha = 2$	Avg. PSNR (dB) ↑			Avg. SSIM ↑			Avg. runtime (sec.) ↓
(Avg. SNR: 33.24dB)	Overall	Natural	Unnatural	Overall	Natural	Unnatural	
HIO [19]	19.79	19.73	21.92	0.50	0.50	0.49	0.25
prDeep [10]	23.45	23.49	21.72	0.65	0.66	0.58	59.32
DIR [33]	23.61	23.60	24.02	0.72	0.72	0.73	21.59
MBwDDP [34]	24.87	24.86	25.56	0.74	0.74	0.74	24.11
Initialization procedure	21.12	21.02	24.82	0.55	0.55	0.58	0.91
InDI-PR ($T = 4$, no ensembling)	28.59	28.65	26.39	0.79	0.80	0.67	1.10
$\alpha = 3$	Avg. PSNR (dB) ↑			Avg. SSIM ↑			Avg. runtime (sec.) ↓
(Avg. SNR: 31.53dB)	Overall	Natural	Unnatural	Overall	Natural	Unnatural	
HIO [19]	18.92	18.89	20.34	0.43	0.43	0.43	0.27
prDeep [10]	22.06	22.09	20.91	0.59	0.59	0.54	59.41
DIR [33]	22.87	22.85	23.50	0.68	0.68	0.71	21.72
MBwDDP [34]	23.92	23.92	23.98	0.70	0.70	0.69	24.35
Initialization procedure	20.17	20.12	22.09	0.51	0.51	0.54	0.90
InDI-PR ($T = 4$, no ensembling)	26.78	26.85	24.18	0.73	0.73	0.61	1.11
$\alpha = 4$	Avg. PSNR (dB) ↑			Avg. SSIM ↑			Avg. runtime (sec.) ↓
(Avg. SNR: 30.24dB)	Overall	Natural	Unnatural	Overall	Natural	Unnatural	
HIO [19]	18.52	18.48	19.80	0.39	0.39	0.40	0.28
prDeep [10]	20.69	20.70	20.38	0.53	0.53	0.51	59.68
DIR [33]	21.80	21.77	22.79	0.62	0.62	0.69	21.95
MBwDDP [34]	22.41	22.39	23.09	0.63	0.63	0.65	24.43
Initialization procedure	19.54	19.46	20.88	0.48	0.48	0.47	0.91
InDI-PR ($T = 4$, no ensembling)	25.43	25.46	24.23	0.66	0.66	0.59	1.10

3.4.2 Effect of Iteration Count

For the evaluation of reconstruction performance for 236 test images under specific settings ($\alpha = 3$ and $2p = 1$), both distortion metrics, such as PSNR and SSIM, and perceptual quality metrics, such as FID [88], LPIPS [89], and CLIP-IQA [90], were utilized. This combination of metrics ensures a comprehensive analysis of image quality from various perspectives, addressing not only the accuracy of pixel values but also the perceptual similarity to human vision.

Table 3.2 shows that a smaller number of iterations tends to yield better outcomes in terms of image quality. This was evident across both types of metrics, suggesting not only higher accuracy and structural fidelity but also greater perceptual similarity to the original images. Additionally, the computational efficiency is enhanced with fewer iterations, as evidenced by faster processing times. However, it is noteworthy that while fewer iterations result in higher metric scores and efficiency, visual inspection of the outputs indicates that larger iteration counts can produce a more varied range of reconstructed images for the same input, suggesting a potential tradeoff between the diversity of output and quantitative performance metrics.

Table 3.2: Average reconstruction performances illustrating the effect of the iteration count for 236 test images with $\alpha = 3$ and no ensembling across 5 Monte Carlo runs.

InDI Total	Perceptual			Distortion		
Iteration Count (T)	FID ↓	LPIPS ↓	CLIP-IQA ↑	PSNR ↑	SSIM ↑	Avg. runtime (sec.) ↓
4	100.96	0.20	0.77	26.78	0.73	1.11
8	103.65	0.21	0.77	26.51	0.71	1.28
32	109.36	0.22	0.77	26.08	0.68	2.41

3.4.3 Effect of Ensembling

Table 3.3 demonstrates the effectiveness of ensembling in image reconstruction, with improvements observed in both perceptual and distortion metrics as the number of different reconstructions increases. This simultaneous enhancement across different quality dimensions suggests that the ensembling approach does not conform to the typical constraints of the perception-distortion tradeoff space, where improvements

in one metric are often countered by compromises in another. Instead, these results indicate that the method is not operating within a Pareto optimal region of this tradeoff space; enhancements are achieved in both perceptual and distortion qualities without the expected tradeoffs.

The success of ensembling in improving these metrics can be attributed to its ability to integrate multiple reconstructions into a single output, effectively averaging out errors and anomalies specific to individual outputs. This process not only increases the overall fidelity and structural integrity of the final image but also preserves the best features of each reconstruction while reducing the impact of any individual output's weaknesses.

However, the computational demands increase with the number of reconstructions, reflecting a significant tradeoff between improved image quality and processing efficiency, which is particularly important in scenarios where speed is crucial.

Table 3.3: Average reconstruction performances showing the effect of the ensembling for 236 test images under $\alpha = 3$ and $T = 32$ setting (5 Monte Carlo runs).

Number of Different	Perceptual			Distortion		
Reconstructions ($2p$)	FID ↓	LPIPS ↓	CLIP-IQA ↑	PSNR ↑	SSIM ↑	Avg. runtime (sec.) ↓
no ensembling	109.36	0.22	0.77	26.08	0.68	2.41
4	96.54	0.17	0.76	27.59	0.77	10.24
6	95.01	0.16	0.76	27.83	0.79	13.55
8	94.26	0.16	0.76	27.95	0.80	20.73
12	93.64	0.15	0.76	28.18	0.81	29.24
24	93.12	0.15	0.76	28.37	0.82	56.51

3.4.4 Uncertainty Quantification Properties

The uncertainty quantification in image reconstruction is crucial for assessing the reliability of the outputs, particularly in scenarios where decisions are based on these images. By utilizing the variance across an ensemble of generated outputs, we estimate the uncertainty in our final reconstruction. The ensemble's variance provides a robust approximation of the expected squared error between the true image $\mathbf{x}$ and the

final reconstructed image $\hat{\mathbf{x}}_{\text{final}}$, as illustrated in the formula:

$$\mathbb{E}\{\|\mathbf{x} - \hat{\mathbf{x}}_{\text{final}}\|^2\} \approx \mathbb{E}\{\|\bar{\mathbf{x}}_{\text{final}} - \hat{\mathbf{x}}_{\text{final}}\|^2\} \approx \text{Var}\{\hat{\mathbf{x}}_{\text{final}}^{(q)}\}_{q=1}^{2p}. \tag{3.6}$$

This statistical approach leverages the diversity within the ensemble to reflect uncertainty, capturing variations that might not be evident when considering a single output.

In our experiments, we also tried to use $\text{Var}\{\hat{\mathbf{x}}_{\text{final}}^{(q)}\}_{q=1}^{2p}$ for the error estimate of the ensemble average output.

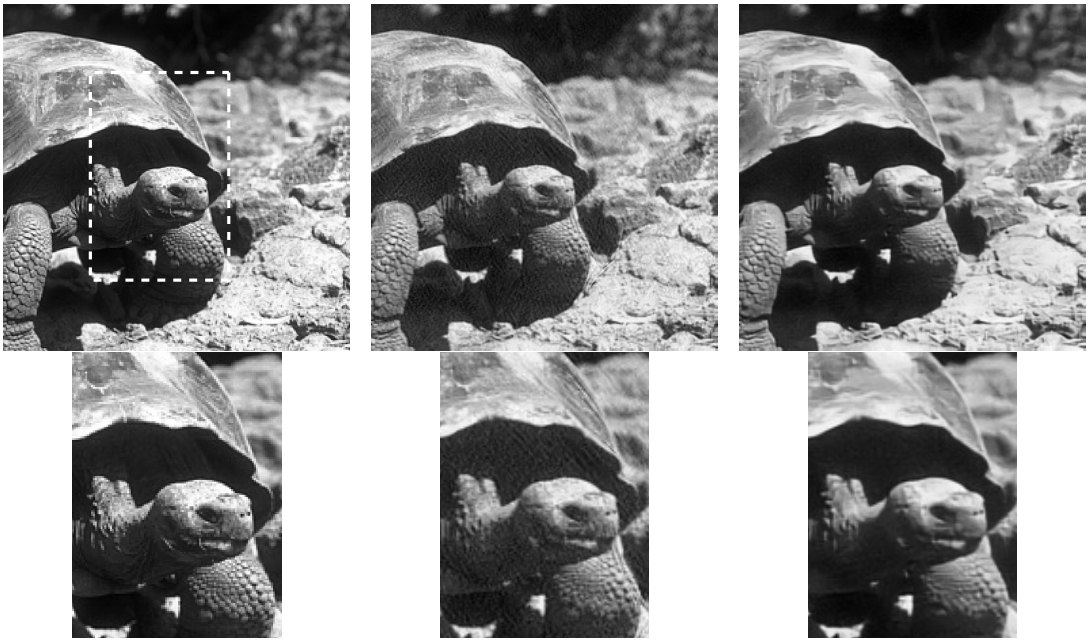
Isotonic regression is used for calibration due to its superior ability to address monotonic distortions by fitting a nondecreasing function to the data, thereby enhancing the calibration accuracy of probabilistic predictions [91]. This calibration is crucial for aligning our model's confidence with the actual performance, as uncalibrated predictions can mislead the decision-making process based on these images.

Fig. 3.9 shows the calibration curves for both uncalibrated and calibrated cases for $2p = 24$, illustrating the improvement in empirical coverage after calibration. The ideal coverage line serves as a benchmark for perfect calibration, highlighting the effectiveness of isotonic regression in aligning our probabilistic predictions with empirical outcomes. Moreover, Fig. 3.10 displays the actual errors and predicted uncertainties before and after calibration for $2p = 24$.

(a) Ground truth

(b) prDeep, PSNR:25.35, SSIM:0.71

(c) DIR, PSNR:26.49, SSIM:0.73



(d) Initialization procedure, PSNR:21.62, SSIM:0.50

(e) InDI-PR ($T = 4, 2p = 1$), PSNR:27.07, SSIM:0.77

(f) InDI-PR ($T = 32, 2p = 24$), PSNR:27.13, SSIM:0.86

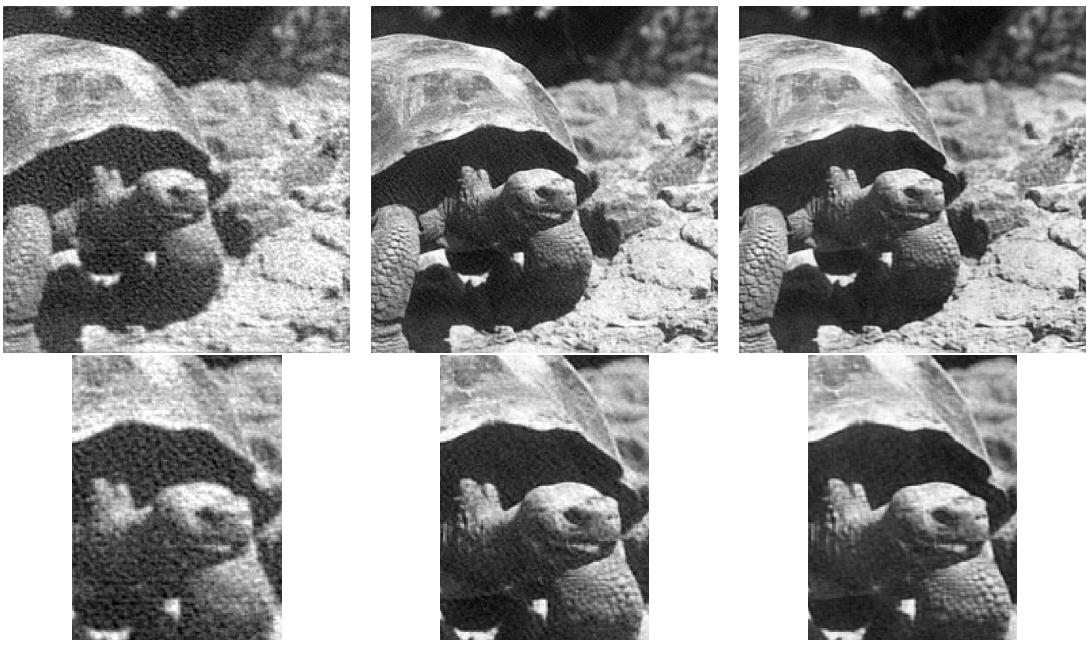


Figure 3.6: The outputs of various algorithms for the "Turtle" test image subjected to $\alpha = 3$ noise (SNR=31.89dB).

(a) Ground truth

(b) prDeep, PSNR:23.44, SSIM:0.66

(c) DIR, PSNR:20.37, SSIM:0.54

(d) Initialization procedure, PSNR:18.96, SSIM:0.33

(e) InDI-PR ($T = 4, 2p = 1$), PSNR:27.39, SSIM:0.70

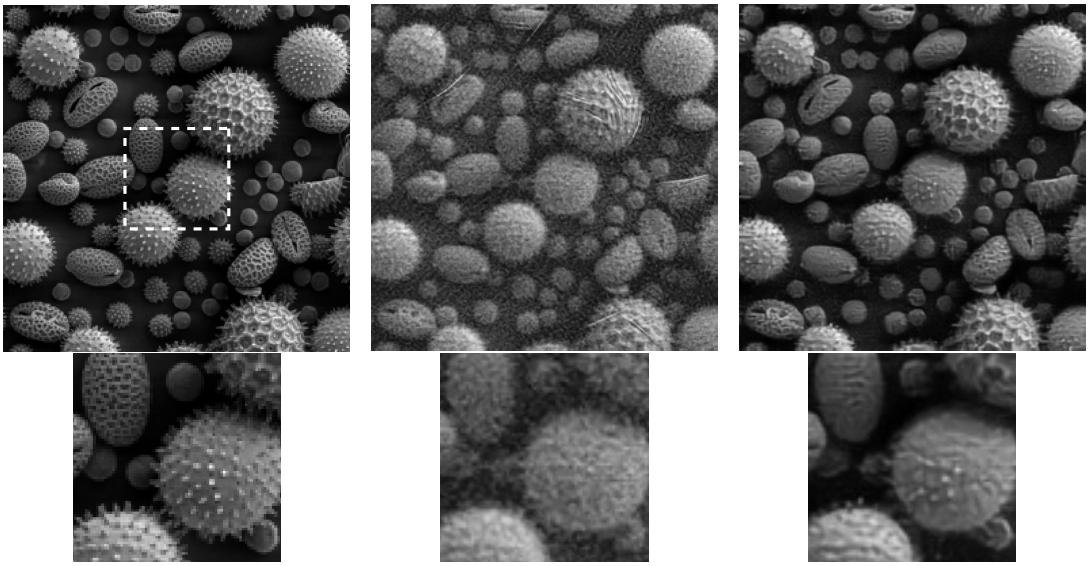
(f) InDI-PR ($T = 32, 2p = 24$), PSNR:28.98, SSIM:0.82

Figure 3.7: The outputs of various algorithms for the "Cameraman" test image subjected to $\alpha = 3$ noise (SNR=31.61dB).

(a) Ground truth

(b) prDeep, PSNR:19.37, SSIM:0.47

(c) DIR, PSNR:25.33, SSIM:0.67



(d) MBwDDP, PSNR:26.28, SSIM:0.71

(e) InDI-PR ($T = 4, 2p = 1$), PSNR:29.10, SSIM:0.87

(f) InDI-PR ($T = 32, 2p = 24$), PSNR:30.46, SSIM:0.90

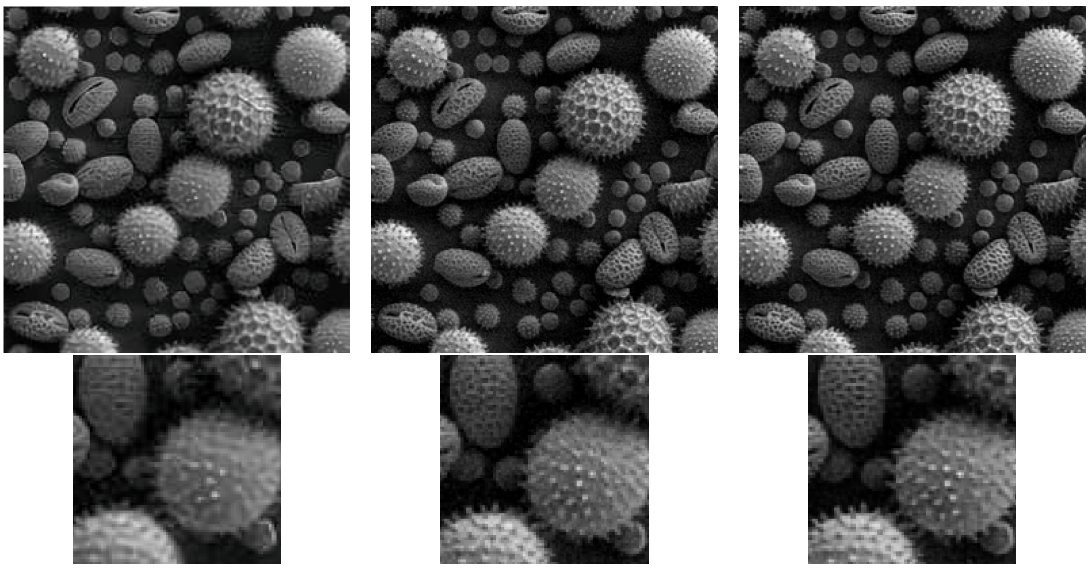


Figure 3.8: The outputs of various algorithms for the out-of-domain "Pollen" test image subjected to $\alpha = 3$ noise (SNR=28.10dB).

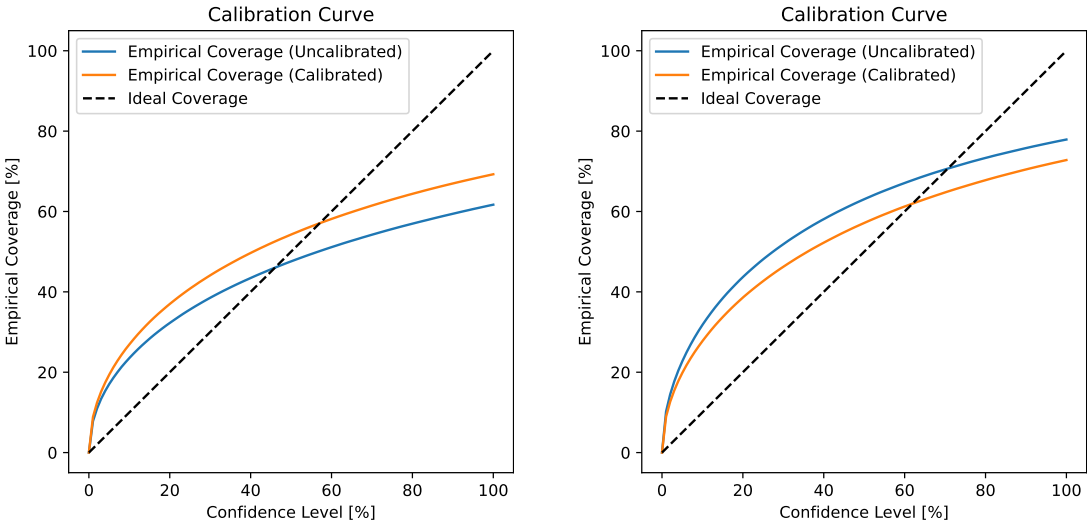


Figure 3.9: Calibration curves for two different cases: for only one output of the algorithm (left), the ensemble average of many output samples (right).

(a) Actual error (log)

(b) Predicted uncertainties (before calibration)

(c) Predicted uncertainties (after calibration)

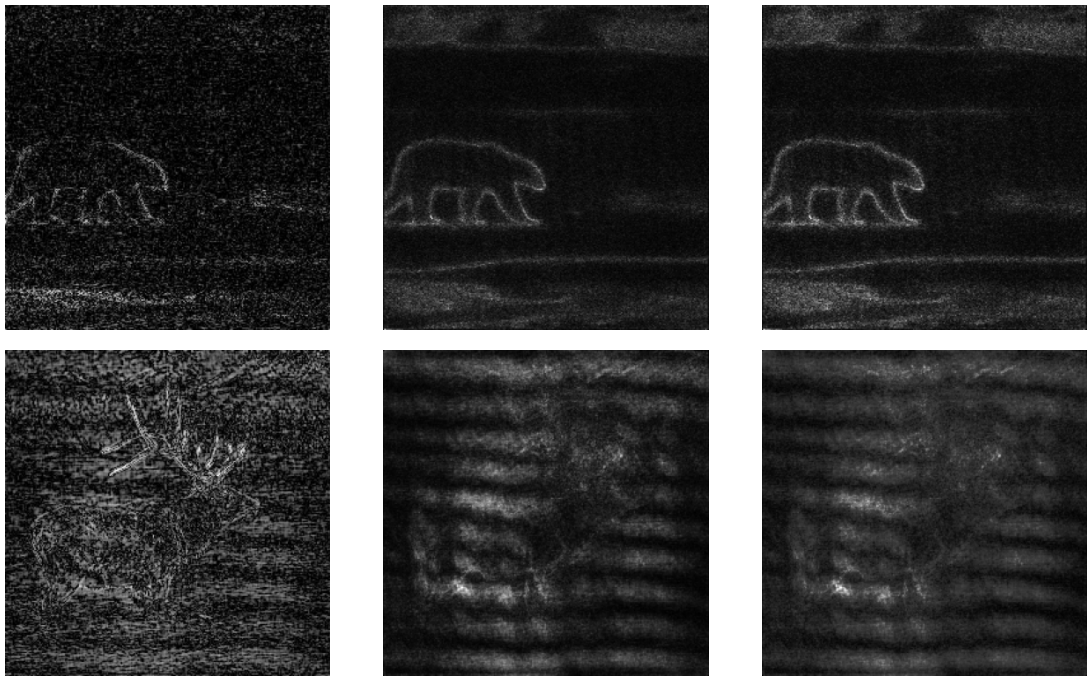


Figure 3.10: Example uncertainty predictions and actual errors for the ensemble average of many output samples.

3.5 Conclusion

This chapter introduces a novel approach to Fourier phase retrieval by employing the Inversion by Direct Denoising (InDI) framework, marking a significant enhancement over traditional methods that commonly initiate from random noise. Our methodology incorporates a sophisticated initialization strategy, utilizes ensembling to refine PSNR metrics, and effectively adapts the InDI process for phase retrieval, showcasing substantial improvements in both training efficiency and image quality.

The implementation of the InDI approach in phase retrieval offers a remarkable improvement by leveraging initial estimates more efficiently. This strategy not only expedites the training process but also ensures that it is more focused on refining rather than reconstructing from scratch, thus enhancing both image quality and computational efficiency.

Moreover, our adoption of ensembling techniques has been shown to enhance both perceptual and distortion metrics simultaneously. This result is indicative of the method's capacity to yield improvements in image reconstruction quality without adhering strictly to the typical constraints of the perception-distortion tradeoff space. This observation suggests that while our method advances current capabilities, it also highlights the potential for further optimization and refinement to achieve even closer approximation to the Pareto optimal frontier in future work.

Our contributions significantly extend the scope of methodological advancements, providing a robust framework that adeptly handles diverse imaging conditions and varying levels of noise. The comprehensive evaluation against established methods confirms our approach's superior performance in terms of reconstruction accuracy and efficiency.

In summary, the techniques developed herein advance the field of classical Fourier phase retrieval and indicate promising avenues for application to other types of phase retrieval challenges. By integrating advanced denoising strategies with novel initialization and ensembling techniques within the InDI framework, this work paves the way for more accurate and efficient phase retrieval methods, potentially enhancing a variety of scientific and industrial applications.

CHAPTER 4

DDRM-PR: FOURIER PHASE RETRIEVAL USING DENOISING DIFFUSION RESTORATION MODELS

4.1 Introduction

In recent years, deep learning has revolutionized the approach to solving inverse problems in imaging, including phase retrieval. Deep neural networks (DNNs) have achieved significant success in directly reconstructing images from measurements or enhancing initial estimates from classical methods. Model-based optimization schemes have also integrated deep priors within the plug-and-play framework. Nevertheless, existing deep learning solutions for PR are often hindered by domain shifts, lack of interpretability, and the necessity for extensive training [24].

Diffusion models have revolutionized the field of unconditional image generation, demonstrating superior performance across various tasks such as super-resolution, deblurring, inpainting, colorization, and compressive sensing. These models gradually and stochastically denoise a sample to produce the desired output, conditioned on the measurements and the inverse problem. The use of pretrained diffusion models allows for efficient and effective restoration without the need for specific training on individual degradation models, thereby offering great flexibility and adaptability in real-world applications [55].

In this work, we extend the efficient, unsupervised posterior sampling method of Denoising Diffusion Restoration Models (DDRM) to the nonlinear inverse problem of phase retrieval. Unlike existing methods, our approach does not require training; instead, it utilizes a pretrained unconditional diffusion model akin to plug-and-play methods. This characteristic significantly enhances the practicality and ease of im-

plementation, as it eliminates the need for additional training and complex parameter tuning [53].

The main contributions of this chapter are as follows:

- We present an innovative method that adapts the DDRM framework to the nonlinear inverse problem of phase retrieval, leveraging pretrained unconditional diffusion models.

- Our approach combines state-of-the-art generative diffusion models with the model-based Hybrid Input-Output (HIO) method, enhancing reconstruction quality.

- We demonstrate the superior performance of our method through empirical evaluations using distortion and perceptual quality metrics between ground truth and reconstructed images, highlighting its potential to outperform classical iterative techniques in phase retrieval.

By integrating the strengths of pretrained diffusion models with classical optimization techniques, our method provides a robust and efficient solution to the challenging problem of phase retrieval, paving the way for further advancements in this field. The developed method is highly versatile and can be easily extended to other types of phase retrieval problems beyond classical Fourier PR, such as coded diffraction pattern (CDP) phase retrieval.

The following sections of this chapter are organized as follows: Section 4.2 reviews related research. Our proposed approach is detailed in Section 4.3, followed by a comparative performance analysis against classical and state-of-the-art methods in Section 4.4. Lastly, Section 4.5 summarizes our findings and outlines future research directions.

4.2 Related Works

4.2.1 Diffusion Models

Diffusion models possess a Markov chain structure, represented as $\mathbf{x}_T \rightarrow \mathbf{x}_{T-1} \rightarrow \ldots \rightarrow \mathbf{x}_1 \rightarrow \mathbf{x}_0$, where $\mathbf{x}_t \in \mathbb{R}^n$. This structure defines their joint distribution as follows:

$$p_\theta\left(\mathbf{x}_{0:T}\right) = p_\theta^{(T)}\left(\mathbf{x}_T\right) \prod_{t=0}^{T-1} p_\theta^{(t)}\left(\mathbf{x}_t \mid \mathbf{x}_{t+1}\right) \tag{4.1}$$

After generating $\mathbf{x}_{0:T}$, only $\mathbf{x}_0$ is retained as the sample from the generative model. To train a diffusion model, a fixed, factorized variational inference distribution is introduced:

$$q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right) = q^{(T)}\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \prod_{t=0}^{T-1} q^{(t)}\left(\mathbf{x}_t \mid \mathbf{x}_{t+1}, \mathbf{x}_0\right) \tag{4.2}$$

This approach results in an evidence lower bound (ELBO) on the maximum likelihood objective. Certain diffusion models have the unique characteristic where both $p_\theta^{(t)}$ and $q^{(t)}$ are defined as conditional Gaussian distributions for all $t < T$. Additionally, $q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)$ is a Gaussian distribution with known mean and covariance, enabling $\mathbf{x}_t$ to be viewed as $\mathbf{x}_0$ corrupted by Gaussian noise. Mathematically, this is expressed as $q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) = \mathcal{N}\left(\sqrt{\alpha_t}\mathbf{x}_0, \left(1-\alpha_t\right)\boldsymbol{I}\right), \quad \forall t \in [1, T]$. As a result, the ELBO objective simplifies into the denoising autoencoder objective, as detailed in [55]:

$$\sum_{t=1}^{T} \gamma_t \mathbb{E}_{\left(\mathbf{x}_0, \mathbf{x}_t\right) \sim q\left(\mathbf{x}_0\right) q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)} \left[\left\| \mathbf{x}_0 - f_\theta^{(t)}\left(\mathbf{x}_t\right) \right\|_2^2 \right] \tag{4.3}$$

Here, $f_\theta^{(t)}$ represents a neural network parameterized by θ, which aims to recover a noiseless observation from a noisy $\mathbf{x}_t$. Additionally, $\gamma_{1:T}$ denotes a set of positive coefficients dependent on $q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)$.

4.2.1.1 Denoising Diffusion Restoration Models (DDRM)

DDRM have emerged as a versatile solution for addressing linear inverse problems in both noisy and noiseless contexts, i.e., $\mathbf{y} = \boldsymbol{H}\mathbf{x} + \mathbf{z}$ where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{y}}^2\mathbf{I})$. Specifically, DDRM functions as a general solver for these problems, defined by the proba-

bilistic model:

$$p_\theta\left(\mathbf{x}_{0:T} \mid \mathbf{y}\right)=p_\theta^{(T)}\left(\mathbf{x}_T \mid \mathbf{y}\right) \prod_{t=0}^{T-1} p_\theta^{(t)}\left(\mathbf{x}_t \mid \mathbf{x}_{t+1}, \mathbf{y}\right), \tag{4.4}$$

where $\mathbf{x}_0$ represents the final diffusion output. In short, the core concept of DDRM is to utilize the singular value decomposition (SVD) of the matrix $\boldsymbol{H}$, transforming both the target variable x, and, the potentially noisy observations y into a common spectral space. Within this spectral space, DDRM distinguishes between dimensions based on the availability of information from y, as indicated by the singular values. For dimensions corresponding to non-zero singular values, DDRM performs denoising, while for those associated with zero singular values, it undertakes imputation. This approach explicitly accounts for measurement noise, thereby enhancing the robustness and accuracy of the restoration process [53].

DDRM employs a procedure that leverages a pretrained unconditional diffusion model to solve various linear inverse problems, akin to plug-and-play methods. Notably, this approach eliminates the necessity for additional training. The authors demonstrate that, under specific conditions, the solution obtained by training a conditional diffusion model is equivalent to that derived from using a pretrained unconditional diffusion model in conjunction with the DDRM procedure. Consequently, this equivalence implies that one can effectively address any linear inverse problem by utilizing a pretrained unconditional model, thus simplifying the implementation and enhancing the practicality of the method. The detailed DDRM procedure can be seen in Appendix C.

4.3 Developed Method

For the case of no noise in the observation $\mathbf{y}$, the general DDRM procedure for linear inverse problems simplifies to be

$$\begin{aligned}
\mathbf{x}_t' &= f_\theta^{(t+1)}\left(\mathbf{x}_{t+1}\right)-\boldsymbol{H}^\dagger \boldsymbol{H} f_\theta^{(t+1)}\left(\mathbf{x}_{t+1}\right)+\boldsymbol{H}^\dagger \mathbf{y} \\
\mathbf{x}_t &= \sqrt{\alpha_t}\left(\eta_b \mathbf{x}_t'+\left(1-\eta_b\right) f_\theta^{(t+1)}\left(\mathbf{x}_{t+1}\right)\right)+\sqrt{1-\alpha_t}\left(\eta \epsilon_t+(1-\eta) \epsilon_\theta^{(t+1)}\left(\mathbf{x}_{t+1}\right)\right)
\end{aligned} \tag{4.5}$$

where $\boldsymbol{H}^\dagger$ represents the Moore-Penrose pseudo-inverse of $\boldsymbol{H}$. The term $f_\theta^{(t+1)}\left(\mathbf{x}_{t+1}\right)$ corresponds to the output of the denoising model at step $t+1$, and $\epsilon_\theta^{(t+1)}\left(\mathbf{x}_{t+1}\right)=$

$\frac{\mathbf{x}_{t+1}-\sqrt{\alpha_{t+1}}f_\theta^{(t+1)}(\mathbf{x}_{t+1})}{\sqrt{1-\alpha_{t+1}}}$ denotes the predicted noise value [92]. The parameters η and η_b are defined by the user, and $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is a vector drawn from a standard Gaussian distribution (refer to Appendix C for the proof).

DDRM is derived for a linear operator $\boldsymbol{H}$. But, in this chapter, it is extended to nonlinear phase retrieval by using the HIO algorithm as $\boldsymbol{H}^\dagger$.

For a linear operator $\boldsymbol{H}$, its pseudo-inverse, $\boldsymbol{H}^\dagger$, exhibits two key properties:

- $\boldsymbol{H}\boldsymbol{H}^\dagger\boldsymbol{H} = \boldsymbol{H}$, meaning that applying the pseudo-inverse does not alter the original measurement.
- $\boldsymbol{H}^\dagger\boldsymbol{H}\mathbf{x}$ approximates $\mathbf{x}$ closely, providing a least-squares solution.

These properties can be extended to certain nonlinear operators. For example, defining $\boldsymbol{H}$ as the forward operator of the phase retrieval problem, i.e., $\boldsymbol{H}\mathbf{x} = |\mathbf{F}\mathbf{x}|$, the HIO algorithm satisfies similar properties:

- Computing the Fourier magnitudes again after applying the HIO algorithm yields the same measurements.
- The HIO method generally preserves visual similarity, thus applying HIO after computing the Fourier magnitudes generates an image "close" to the original one.

For the phase retrieval problem, the following method is developed for this method.

$$\begin{aligned}\mathbf{x}_t' &= f_\theta^{(t+1)}(\mathbf{x}_{t+1}) - \text{HIO}(|\mathbf{F}f_\theta^{(t+1)}(\mathbf{x}_{t+1})|) + \text{RandomInit}(\mathbf{y}) \\ \mathbf{x}_t &= \sqrt{\alpha_t}\left(\eta_b\mathbf{x}_t' + (1-\eta_b)f_\theta^{(t+1)}(\mathbf{x}_{t+1})\right) + \sqrt{1-\alpha_t}\left(\eta\epsilon_t + (1-\eta)\epsilon_\theta^{(t+1)}(\mathbf{x}_{t+1})\right)\end{aligned} \tag{4.6}$$

Here, RandomInit refers to the HIO initialization procedure in the prDeep paper [10]. Initially, the HIO method was executed with $m = 50$ different random initializations, each for $s = 50$ iterations. Subsequently, the reconstruction with the smallest residual was selected for an additional HIO run of $n = 1000$ iterations.

For, HIO$(|\mathbf{F}f_\theta^{(t+1)}(\mathbf{x}_{t+1})|)$, the algorithm is applied for $k = 100$ steps and initialized with $\mathbf{x}_{t+1}$.

Furthermore, to ensure consistent performance, we generate $N = 8$ independent outputs for each input and use the averaged image obtained from these outputs.

In order to optimize the hyperparameters, such as η, η_b, uniformly-spaced diffusion steps t, initial timestep T_{init}, and the number of averaged samples N, a simple linear grid search is used.

Our method's integration of pretrained unconditional diffusion models offers several practical advantages. The pretrained models are initially developed on large and diverse image datasets, capturing a wide range of features that are crucial for effective denoising and reconstruction. This integration bypasses the need for retraining specific to the phase retrieval task, making the method more accessible and easier to implement in various settings. The pretrained model acts as a strong prior, facilitating accurate reconstruction by refining noisy inputs iteratively. Additionally, we employ a straightforward grid search to optimize hyperparameters, ensuring that the method is not only effective but also user-friendly.

4.4 Results

In our experiments, we used the CelebA-HQ dataset at a resolution of 256x256 pixels to evaluate the effectiveness of our proposed method. The choice of RGB images is twofold: firstly, RGB images tend to reveal artifacts from the HIO algorithm more clearly, and secondly, the pretrained diffusion models we employed are optimized for RGB images. Each color channel (Red, Green, and Blue) is processed separately by the HIO algorithm, ensuring that color information is preserved and accurately reconstructed.

After applying the HIO algorithm to each channel, we calculated the Peak Signal-to-Noise Ratio (PSNR) values to correct for conjugate inversion ambiguity. This step is crucial for ensuring the accuracy of the phase retrieval process, as it aligns the reconstructed images more closely with the ground truth.

To comprehensively evaluate the performance of our algorithm, we conducted experiments under various noise levels. The primary metrics used to assess the quality of

the reconstructed images were PSNR, Structural Similarity Index (SSIM) [80], and Learned Perceptual Image Patch Similarity (LPIPS) [89]. These metrics provide a well-rounded evaluation of both the pixel-level accuracy and the perceptual quality of the reconstructions.

We applied our algorithm to a diverse set of test images from the CelebA-HQ dataset, introducing varying noise levels to simulate real-world scenarios where measurements are often contaminated with different degrees of noise, i.e., different values of α. For each noise level, we generated multiple reconstructions and computed the average values of PSNR, SSIM, and LPIPS to ensure the robustness and consistency of our method.

Our method demonstrated superior performance across all evaluated metrics compared to traditional phase retrieval techniques such as HIO. The PSNR values indicated that our method effectively suppressed noise while preserving image details. The SSIM scores showed that the structural integrity of the images was well-maintained, and the high LPIPS values confirmed the perceptual quality of the reconstructions.

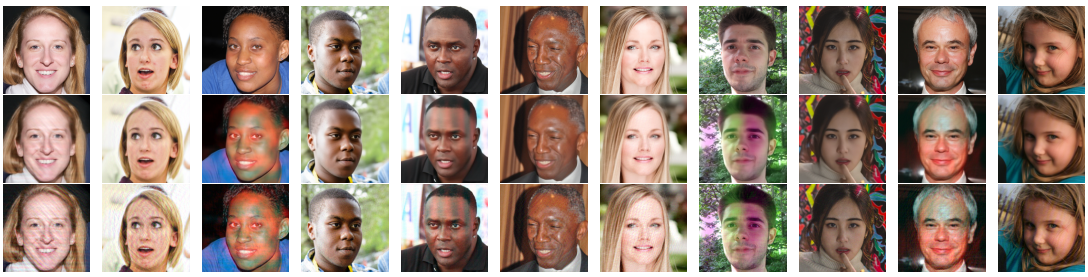


Figure 4.1: Ground-truth test images (top row), reconstructions using the developed approach (middle row), and HIO initialization results (bottom row) for the case with parameters: $\alpha = 0.5$, $N = 1$, $\eta = 0.15$, $\eta_b = 0.20$, $t = 15$, and $T_{init} = 350$.

The high performance of our method can be attributed to the integration of pretrained diffusion models with the HIO algorithm, which allows for effective denoising and accurate phase retrieval. The use of multiple independent outputs and averaging further enhances the stability and reliability of the results. Additionally, the optimization of hyperparameters through a linear grid search ensures that the method is finely tuned for the best possible performance.

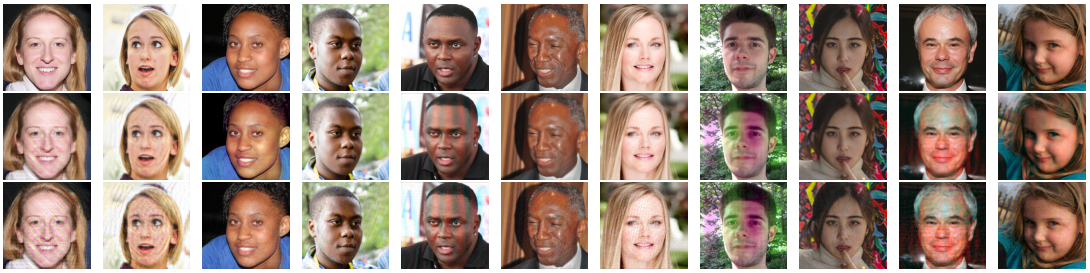


Figure 4.2: Ground-truth test images (top row), reconstructions using the developed approach (middle row), and HIO initialization results (bottom row) for the case with parameters: $\alpha = 1$, $N = 1$, $\eta = 0.25$, $\eta_b = 0.22$, $t = 30$, and $T_{init} = 400$.

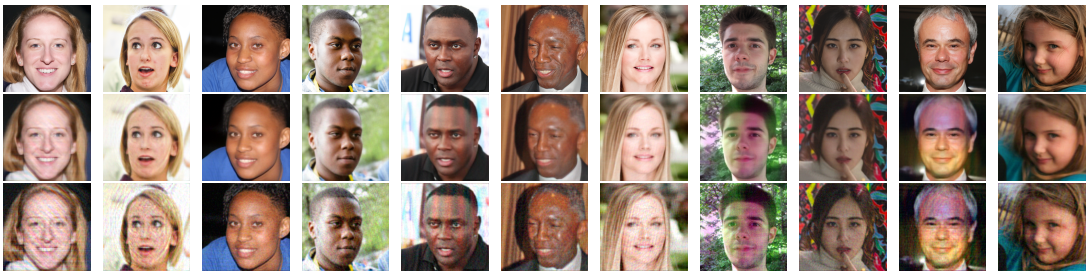


Figure 4.3: Ground-truth test images (top row), reconstructions using the developed approach (middle row), and HIO initialization results (bottom row) for the case with parameters: $\alpha = 2$, $N = 1$, $\eta = 0.25$, $\eta_b = 0.18$, $t = 15$, and $T_{init} = 400$.

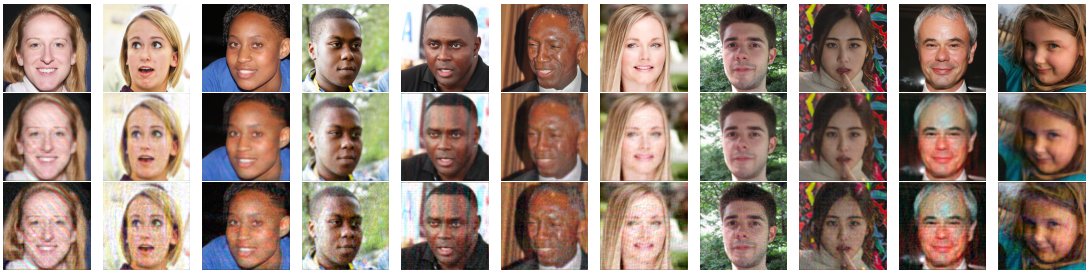


Figure 4.4: Ground-truth test images (top row), reconstructions using the developed approach (middle row), and HIO initialization results (bottom row) for the case with parameters: $\alpha = 3$, $N = 1$, $\eta = 0.78$, $\eta_b = 0.17$, $t = 30$, and $T_{init} = 300$.

In conclusion, the results of our experiments validate the efficacy of our proposed method in solving the nonlinear inverse problem of phase retrieval. The comprehensive evaluation using PSNR, SSIM, and LPIPS metrics confirms that our approach outperforms existing techniques, providing a robust and reliable solution for high-

quality image reconstruction in the presence of noise. As demonstrated qualitatively in Figs. 4.1, 4.2, 4.3, and 4.4, and quantitatively in Table 4.1, our method consistently produces superior results.

Table 4.1: Average reconstruction performances of the developed algorithms for different images from the CelebA-HQ test set.

	$\alpha = 0.5$			$\alpha = 1$			$\alpha = 2$			$\alpha = 3$		
Methods	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
HIO Stage	28.74	0.82	0.14	27.57	0.74	0.21	25.27	0.65	0.34	24.00	0.58	0.43
DDRM-PR	29.13	0.87	0.13	28.45	0.84	0.15	26.59	0.79	0.23	25.73	0.76	0.27

4.5 Conclusion

In this chapter, we present a new approach for addressing the nonlinear inverse problem of phase retrieval by extending the Denoising Diffusion Restoration Models (DDRM) framework. Our method uniquely leverages pretrained unconditional diffusion models, eliminating the need for additional training and aligning with the plug-and-play paradigm. This characteristic significantly enhances the practicality and ease of implementation of our approach.

Through empirical evaluations, we demonstrate the superior performance of our method in phase retrieval tasks. The proposed technique not only achieves high-quality reconstructions but also exhibits robustness across various scenarios. Our results, evaluated using photometric similarity metrics between ground truth and reconstructed images, underscore the efficacy of our approach in overcoming the limitations of existing methods.

A key advantage of our method is its ability to generalize beyond the specific task of phase retrieval without necessitating retraining or extensive problem-specific hyperparameter tuning. This flexibility highlights the potential of our approach to be applied to a broader range of inverse problems in imaging.

In summary, our work contributes a novel, practical solution to the phase retrieval problem by integrating state-of-the-art generative diffusion models with the Hybrid Input-Output (HIO) method.

CHAPTER 5

CONCLUSION

In this thesis, we have developed and evaluated novel data-driven phase retrieval methods that leverage deep learning and diffusion models to address the long-standing challenges of this fundamental problem in optical systems and many other areas. The work presented in this thesis encompasses significant advancements in phase retrieval, focusing on improving reconstruction quality, robustness, and computational efficiency.

Chapter 2 introduced a novel approach to phase retrieval by employing Langevin dynamics for posterior sampling within the framework of score/diffusion-based models. This method, realized through the development of the prNet-Small and prNet-Large pipelines, relies on the iterative refinement of the initial HIO estimates through denoising, data consistency, and noise injection cycles. By paying attention to the perception-distortion tradeoff, the method not only yields high-fidelity reconstructions with low distortion but also achieves high perceptual quality. prNet-Large, in particular, demonstrated enhanced robustness and perceptual quality by incorporating diverse starting points and employing an additional denoiser with a Wasserstein loss. Extensive simulations confirmed the state-of-the-art performance of this approach with low computational cost, indicating that this approach can be extended as a reliable stochastic nonlinear inverse problem solver.

In Chapter 3, we applied the Inversion by Direct Denoising (InDI) framework to the Fourier phase retrieval problem. This novel method utilizes advanced initialization strategies and ensembling techniques to improve quality metrics and enhance training efficiency. By starting from a plausible initial estimate rather than random noise, the InDI framework makes full use of the denoiser's capacity, reducing train-

ing time while demonstrating superior performance compared to both classical and contemporary techniques. This method sets a new benchmark for phase retrieval by significantly improving both training efficiency and image quality.

Chapter 4 extended the application of Denoising Diffusion Restoration Models (DDRM) from linear inverse problems to the nonlinear inverse problem of phase retrieval. By combining state-of-the-art generative diffusion models with the Hybrid Input-Output (HIO) method, we applied pretrained unconditional diffusion models to phase retrieval. The results demonstrated that this combined approach outperforms existing classical iterative methods, providing a powerful tool for phase retrieval without requiring any training.

The integration of deep learning into phase retrieval represents a significant advancement, offering new solutions to long-standing challenges and opening new possibilities for coherent imaging. By learning from large datasets, deep learning models provide robust priors that guide the phase retrieval process, reducing the impact of noise and improving convergence to accurate solutions.

The methods developed in this thesis are based on the score/diffusion-based framework, which has gained prominence for its effectiveness in high-dimensional data generation and reconstruction tasks. The iterative nature of these models aligns well with the needs of phase retrieval, allowing for incremental refinement of solutions and making them well-suited for tasks requiring high precision.

Overall, this thesis has demonstrated that the hybrid use of deep learning models with traditional model-based techniques can significantly enhance the performance of phase retrieval algorithms. The developed methods—prNet-Small, prNet-Large, InDI-PR, and DDRM-PR—offer diverse solutions tailored to different aspects of phase retrieval, showcasing the flexibility and power of combining deep learning with classical approaches.

Future research directions include developing new denoising architectures specifically tailored for phase retrieval and applying these methods to real-world experimental data. Additionally, extending these techniques to other types of phase retrieval problems and nonlinear inverse problems, as well as exploring their integration with

emerging imaging modalities, will further advance the field. The ongoing convergence of deep learning and phase retrieval holds the promise of exciting developments and innovations, enhancing our ability to capture and reconstruct complex signals and images with unprecedented accuracy and detail.

To conclude, this thesis has made substantial contributions to the field of phase retrieval by introducing innovative methods that combine deep learning and generative models with traditional phase retrieval techniques. These advancements not only improve the accuracy and efficiency of phase retrieval but also pave the way for new applications and further research in optical systems and other related fields. The promising results obtained from the developed methods underscore the potential of deep generative learning to revolutionize the approach to solving complex inverse problems in science and engineering.

REFERENCES

[1] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, "Phase retrieval with application to optical imaging: A contemporary overview", *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 87–109, 2015.

[2] J. Dong, L. Valzania, A. Maillard, T.-a. Pham, S. Gigan, and M. Unser, "Phase retrieval: From computational imaging to machine learning: A tutorial", *IEEE Signal Processing Magazine*, vol. 40, no. 1, pp. 45–57, Jan. 2023, ISSN: 1558-0792.

[3] J. R. Fienup, "Phase retrieval algorithms: A personal tour", *Appl. Opt.*, vol. 52, no. 1, pp. 45–56, Jan. 2013.

[4] A. Walther, "The question of phase retrieval in optics", *Optica Acta: International Journal of Optics*, vol. 10, no. 1, pp. 41–49, 1963.

[5] T. J. Schulz and D. L. Snyder, "Image recovery from correlations", *J. Opt. Soc. Am. A*, vol. 9, no. 8, pp. 1266–1272, Aug. 1992.

[6] J. C. Dainty and J. Fienup, "Phase retrieval and image reconstruction for astronomy", *Image Recovery: Theory Appl*, vol. 13, pp. 231–275, Jan. 1987.

[7] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Prentice-Hall, Inc., 1993.

[8] K. Khare, M. Butola, and S. Rajora, "Fourier optics and computational imaging", 2023.

[9] C. Liu, S. Wang, and S. P. Veetil, "Computational optical phase imaging", *Progress in Optical Science and Photonics*, 2022.

[10] C. Metzler, P. Schniter, A. Veeraraghavan, and R. Baraniuk, “PrDeep: Robust phase retrieval with a flexible deep network”, in *Proceedings of the 35th International Conference on Machine Learning*, JMLR.org, 2018, pp. 3498–3507.

[11] M. Hayes, “The reconstruction of a multidimensional sequence from the phase or magnitude of its fourier transform”, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 2, pp. 140–154, Apr. 1982, ISSN: 0096-3518.

[12] J. Miao, P. S. Charalambous, J. Kirz, and D. Sayre, “Extending the methodology of x-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens”, *Nature*, vol. 400, pp. 342–344, 1999.

[13] T. Latychevskaia, “Iterative phase retrieval in coherent diffractive imaging: Practical issues”, *Appl. Opt.*, vol. 57, no. 25, pp. 7187–7197, Sep. 2018.

[14] R. A. Kirian, *Classical far-field elastic x-ray scattering under the first born approximation*, 2024.

[15] J. W. Goodman, “Introduction to fourier optics”, 1969.

[16] Y. Shechtman, A. Szameit, E. Bullkich, *et al.*, “Sparsity-based single-shot subwavelength coherent diffractive imaging”, *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel*, pp. 1–4, 2011.

[17] R. W. Gerchberg and W. O. Saxton, “A practical algorithm for the determination of phase from image and diffraction plane pictures”, *Optik*, vol. 35, pp. 237–250, Nov. 1972.

[18] J. R. Fienup, “Reconstruction of an object from the modulus of its fourier transform”, *Optics letters*, vol. 3, no. 1, pp. 27–29, 1978.

[19] J. R. Fienup, “Phase retrieval algorithms: A comparison”, *Appl. Opt.*, vol. 21, no. 15, pp. 2758–2769, Aug. 1982.

[20] S. Marchesini, “Invited article: A unified evaluation of iterative projection algorithms for phase retrieval”, *Review of scientific instruments*, vol. 78, no. 1, 2007.

[21] J. Qian, C. Yang, A. Schirotzek, F. Maia, and S. Marchesini, "Efficient algorithms for ptychographic phase retrieval, in inverse problems and applications", *Contemp. Math*, vol. 615, pp. 261–280, Jan. 2014.

[22] A. Maiden, D. Johnson, and P. Li, "Further improvements to the ptychographical iterative engine", *Optica*, vol. 4, no. 7, pp. 736–745, Jul. 2017.

[23] S. López-Tapia, R. Molina, and A. Katsaggelos, "Deep learning approaches to inverse problems in imaging: Past, present and future", *Digital Signal Processing*, vol. 119, p. 103 285, Oct. 2021.

[24] K. Wang, L. Song, C. Wang, *et al.*, "On the use of deep learning for phase recovery", *Light: Science & Applications*, vol. 13, no. 1, Jan. 2024, ISSN: 2047-7538.

[25] M. El Helou and S. Susstrunk, "Blind universal bayesian image denoising with gaussian noise level learning", *IEEE Transactions on Image Processing*, vol. 29, pp. 4885–4897, 2020, ISSN: 1941-0042.

[26] Y. Nishizaki, R. Horisaki, K. Kitaguchi, M. Saito, and J. Tanida, "Analysis of non-iterative phase retrieval based on machine learning", *Optical Review*, vol. 27, pp. 136–141, 2020.

[27] T. Uelwer, A. Oberstrass, and S. Harmeling, "Phase retrieval using conditional generative adversarial networks", *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 731–738, 2019.

[28] I. Shevkunov, J. Kilpeläinen, and K. Eguiazarian, "Deep convolutional neural network-based lensless quantitative phase retrieval", in *BiOS*, 2021.

[29] W. Zhang, Y. Wan, Z. Zhuang, and J. Sun, *What is wrong with end-to-end learning for phase retrieval?*, 2024.

[30] K. H. Jin, M. T. McCann, E. Froustey, and M. A. Unser, "Deep convolutional neural network for inverse problems in imaging", *IEEE Transactions on Image Processing*, vol. 26, pp. 4509–4522, 2016.

[31] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, “Plug-and-play priors for model based reconstruction”, in *2013 IEEE Global Conference on Signal and Information Processing*, 2013, pp. 945–948.

[32] Y. Romano, M. Elad, and P. Milanfar, “The little engine that could: Regularization by denoising (red)”, *ArXiv*, vol. abs/1611.02862, 2016.

[33] Ç. Işil, F. S. Oktem, and A. Koç, “Deep iterative reconstruction for phase retrieval”, *Appl. Opt.*, vol. 58, no. 20, pp. 5422–5431, Jul. 2019.

[34] Ç. Işil and F. S. Oktem, “Model-based phase retrieval with deep denoiser prior”, in *Imaging and Applied Optics Congress*, Optica Publishing Group, 2020.

[35] E. J. Cha, C. Lee, M. Jang, and J. C. Ye, “Deepphasecut: Deep relaxation in phase for unsupervised fourier phase retrieval”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 9931–9943, 2020.

[36] Y. Wang, X. Sun, and J. W. Fleischer, “When deep denoising meets iterative phase retrieval”, *ArXiv*, vol. abs/2003.01792, 2020.

[37] H. K. Aggarwal, M. P. Mani, and M. Jacob, “Modl: Model-based deep learning architecture for inverse problems”, *IEEE Transactions on Medical Imaging*, vol. 38, pp. 394–405, 2017.

[38] V. Monga, Y. Li, and Y. C. Eldar, “Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing”, *IEEE Signal Processing Magazine*, vol. 38, pp. 18–44, 2019.

[39] N. Naimipour, S. Khobahi, and M. Soltanalian, “Upr: A model-driven architecture for deep phase retrieval”, *2020 54th Asilomar Conference on Signals, Systems, and Computers*, pp. 205–209, 2020.

[40] M. Deng, A. Goy, K. Arthur, and G. Barbastathis, “Physics embedded deep neural network for phase retrieval under low photon conditions”, *Imaging and Applied Optics 2019 (COSI, IS, MATH, pcAOP)*, 2019.

[41] N. Naimipour, S. Khobahi, and M. Soltanalian, “Unfolded algorithms for deep phase retrieval”, *ArXiv*, vol. abs/2012.11102, 2020.

[42] C.-J. Wang, C.-K. Wen, S.-H. L. Tsai, and S. Jin, "Phase retrieval with learning unfolded expectation consistent signal recovery algorithm", *IEEE Signal Processing Letters*, vol. 27, pp. 780–784, 2020.

[43] A. Liu, X. Fan, Y. Yang, and J. Zhang, "Prista-net: Deep iterative shrinkage thresholding network for coded diffraction patterns phase retrieval", *ArXiv*, 2023.

[44] Y. Blau and T. Michaeli, "The perception-distortion tradeoff", *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6228–6237, 2017.

[45] C. Ledig, L. Theis, F. Huszár, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network", in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 105–114.

[46] K. Tayal, C.-H. Lai, V. Kumar, and J. Sun, "Inverse problems, deep learning, and symmetry breaking", *ArXiv*, vol. abs/2003.09077, 2020.

[47] A. G. Dimakis, "Deep generative models and inverse problems", in *Mathematical Aspects of Deep Learning*, P. Grohs and G. Kutyniok, Eds. Cambridge University Press, 2022, pp. 400–421.

[48] Z. Zhao, J. C. Ye, and Y. Bresler, "Generative models for inverse imaging problems: From mathematical foundations to physics-driven applications", *IEEE Signal Processing Magazine*, vol. 40, no. 1, pp. 148–163, 2023.

[49] J. Gladrow, "Digital phase-only holography using deep conditional generative models", *ArXiv*, vol. abs/1911.00904, 2019.

[50] S. Shoushtari, J.-M. Liu, and U. S. Kamilov, "Dolph: Diffusion models for phase retrieval", *ArXiv*, vol. abs/2211.00529, 2022.

[51] S. H. Chan, "Tutorial on diffusion models for imaging and vision", *arXiv preprint arXiv:2403.18103*, 2024.

[52] M. Delbracio and P. Milanfar, "Inversion by direct iteration: An alternative to denoising diffusion for image restoration", *arXiv preprint arXiv:2303.11435*, 2023.

[53] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models", in *Advances in Neural Information Processing Systems*, 2022.

[54] C. Luo, "Understanding diffusion models: A unified perspective", *ArXiv preprint*, vol. abs/2208.11970, 2022.

[55] X. Li, Y. Ren, X. Jin, *et al.*, "Diffusion models for image restoration and enhancement, a comprehensive survey", *arXiv preprint arXiv:2308.09388*, 2023.

[56] K. Miyasawa *et al.*, "An empirical bayes estimator of the mean of a normal population", *Bull. Inst. Internat. Statist*, vol. 38, no. 181-188, pp. 1–2, 1961.

[57] B. Kawar, G. Vaksman, and M. Elad, "Snips: Solving noisy inverse problems stochastically", in *Neural Information Processing Systems*, 2021.

[58] Z. Zhao, J. C. Ye, and Y. Bresler, "Generative models for inverse imaging problems: From mathematical foundations to physics-driven applications", *IEEE Signal Processing Magazine*, vol. 40, pp. 148–163, 2023.

[59] P. Hand, O. Leong, and V. Voroninski, "Phase retrieval under a generative prior", in *Neural Information Processing Systems*, 2018.

[60] S. Peng and K. Li, *Generating unobserved alternatives: A case study through super-resolution and decompression*, Nov. 2020.

[61] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan", *ArXiv preprint*, vol. abs/1701.07875, 2017.

[62] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans", in *Neural Information Processing Systems*, 2017.

[63] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning", *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[64] M. Kimura, "Understanding test-time augmentation", in *International Conference on Neural Information Processing*, Springer, 2021, pp. 558–569.

[65] D. Shanmugam, D. Blalock, G. Balakrishnan, and J. Guttag, “Better aggregation in test-time augmentation”, in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1214–1223.

[66] Á. Casado-García and J. Heras, “Ensemble methods for object detection”, in *ECAI 2020*, IOS Press, 2020, pp. 2688–2695.

[67] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, “Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks”, *Neurocomputing*, vol. 338, pp. 34–45, 2019.

[68] M. Delbracio and P. Milanfar, “Inversion by direct iteration: An alternative to denoising diffusion for image restoration”, *ArXiv*, vol. abs/2303.11435, 2023.

[69] C. Saharia, W. Chan, H. Chang, *et al.*, “Palette: Image-to-image diffusion models”, *ACM SIGGRAPH 2022 Conference Proceedings*, 2021.

[70] A. Bansal, E. Borgnia, H.-M. Chu, *et al.*, “Cold diffusion: Inverting arbitrary image transforms without noise”, *ArXiv*, vol. abs/2208.09392, 2022.

[71] J. Whang, M. Delbracio, H. Talebi, C. Saharia, A. G. Dimakis, and P. Milanfar, “Deblurring via stochastic refinement”, *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16 272–16 282, 2021.

[72] D. Martin, C. Fowlkes, D. Tal, and J. Malik, “A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics”, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, IEEE, vol. 2, 2001, pp. 416–423.

[73] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database”, in *Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2009, pp. 248–255.

[74] K. Zhang, W. Zuo, S. Gu, and L. Zhang, “Learning deep CNN denoiser prior for image restoration”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2017, pp. 3929–3938.

[75] K. Ma, Z. Duanmu, Q. Wu, *et al.*, "Waterloo Exploration Database: New challenges for image quality assessment models", *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 1004–1016, Feb. 2017.

[76] R. W. Franzen, "True color kodak images", *http://r0k.us/graphics/kodak*,

[77] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.

[78] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization", in *International Conference on Learning Representations*, 2017.

[79] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts", *arXiv: Learning*, 2016.

[80] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity", *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[81] A. Goy, K. Arthur, S. Li, and G. Barbastathis, "Low photon count phase retrieval using deep learning.", *Physical review letters*, vol. 121 24, p. 243 902, 2018.

[82] S. Marchesini, H. He, H. Chapman, *et al.*, "X-ray image reconstruction from a diffraction pattern alone", *Physical Review B*, vol. 68, Jul. 2003.

[83] F. J. A. Artacho, R. Campoy, and M. K. Tam, "The douglas–rachford algorithm for convex and nonconvex feasibility problems", *Mathematical Methods of Operations Research*, vol. 91, pp. 201–240, 2019.

[84] R. Heckel, M. Jacob, A. Chaudhari, O. Perlman, and E. Shimron, "Deep learning for accelerated and robust mri reconstruction: A review", *arXiv preprint*, vol. abs/2404.15692, 2024.

[85] G.-H. Liu, A. Vahdat, D.-A. Huang, E. A. Theodorou, W. Nie, and A. Anandkumar, "I2sb: Image-to-image schrödinger bridge", in *International Conference on Machine Learning*, 2023.

[86] H. Chung, J. Kim, and J. C. Ye, "Direct diffusion bridge using data consistency for inverse problems", *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[87] M. U. Mirza, O. Dalmaz, H. A. Bedel, *et al.*, "Learning fourier-constrained diffusion bridges for mri reconstruction", 2023.

[88] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a nash equilibrium", *ArXiv*, vol. abs/1706.08500, 2017.

[89] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric", *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

[90] J. Wang, K. C. K. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images", in *AAAI Conference on Artificial Intelligence*, 2022.

[91] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning", *Proceedings of the 22nd international conference on Machine learning*, 2005.

[92] B. Kawar, J. Song, S. Ermon, and M. Elad, "Jpeg artifact correction using denoising diffusion restoration models", *ArXiv*, 2022.

APPENDIX A

PRNET EXAMPLE RECONSTRUCTIONS

This appendix showcases example reconstructions obtained using the proposed methods (prNet-Small and prNet-Large) for various test images.

Figure A.1: The reconstructions of the different algorithms for different test images under the $\alpha = 3$ noise level.

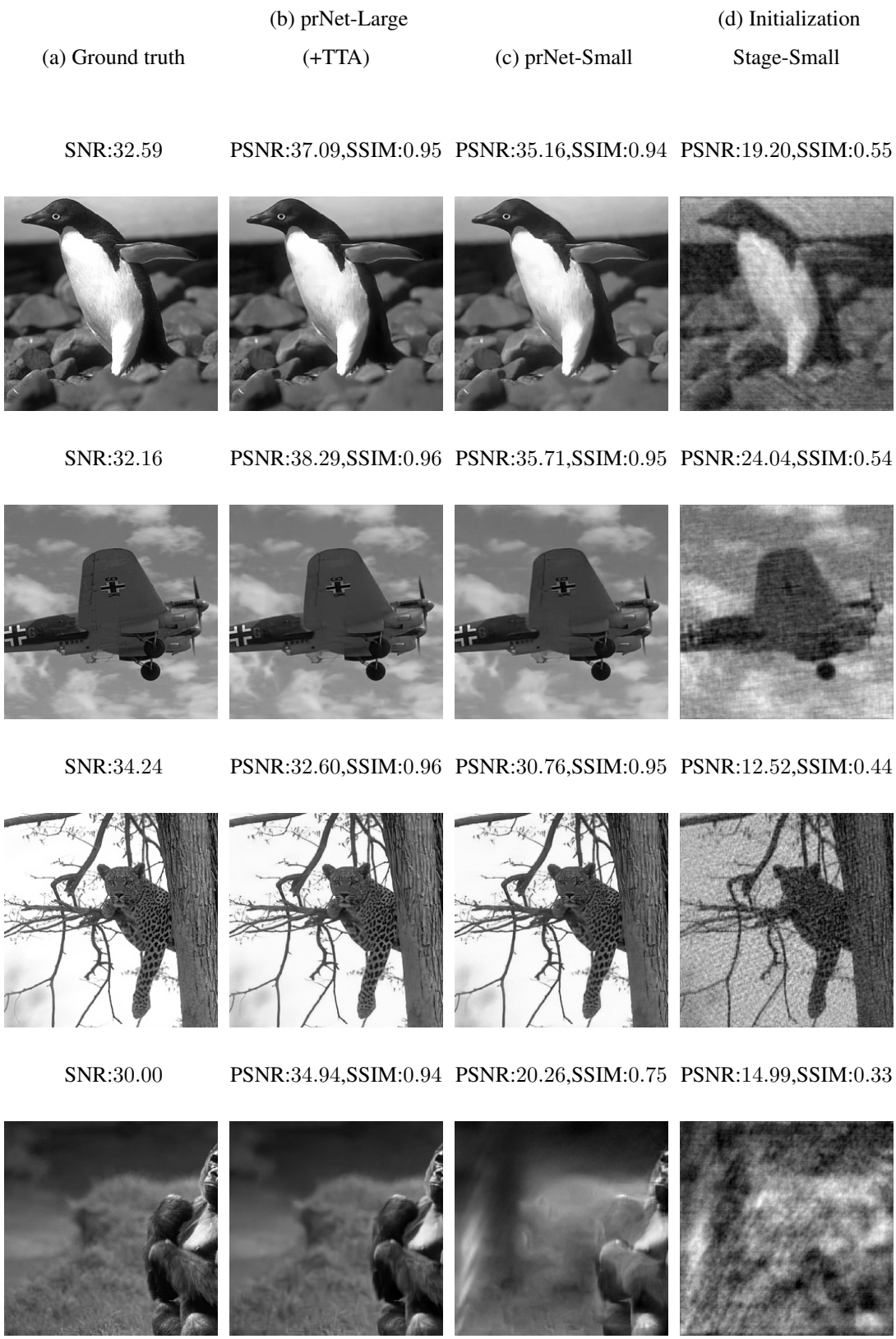


Figure A.2: The outputs of various algorithms for different test set images subjected to $\alpha = 3$ noise.

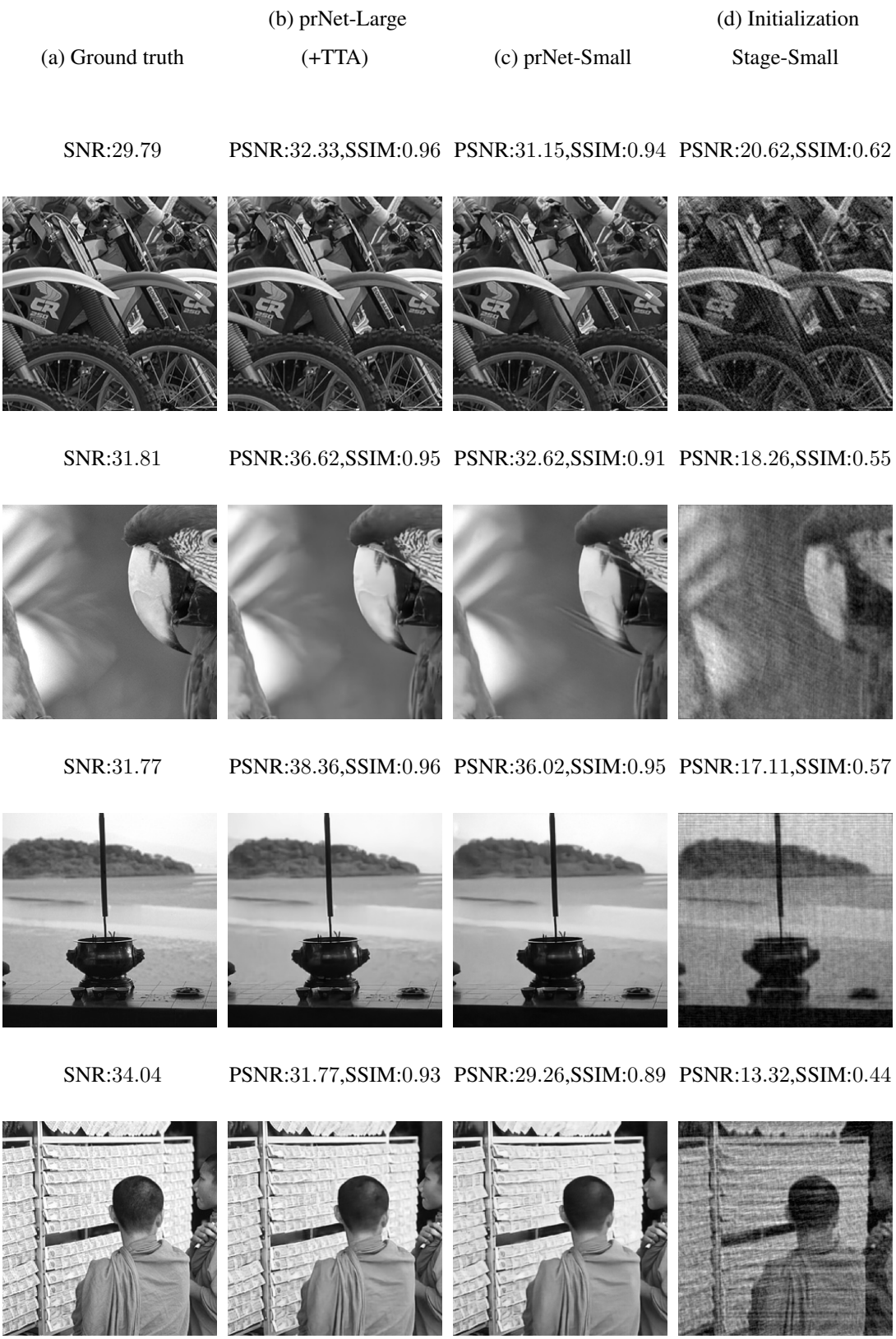


Figure A.3: The outputs of various algorithms for different test set images subjected to $\alpha = 3$ noise.

APPENDIX B

INDI-PR EXAMPLE RECONSTRUCTIONS

This appendix showcases example reconstructions obtained using the proposed methods (InDI-PR and its initialization procedure) for various test images.

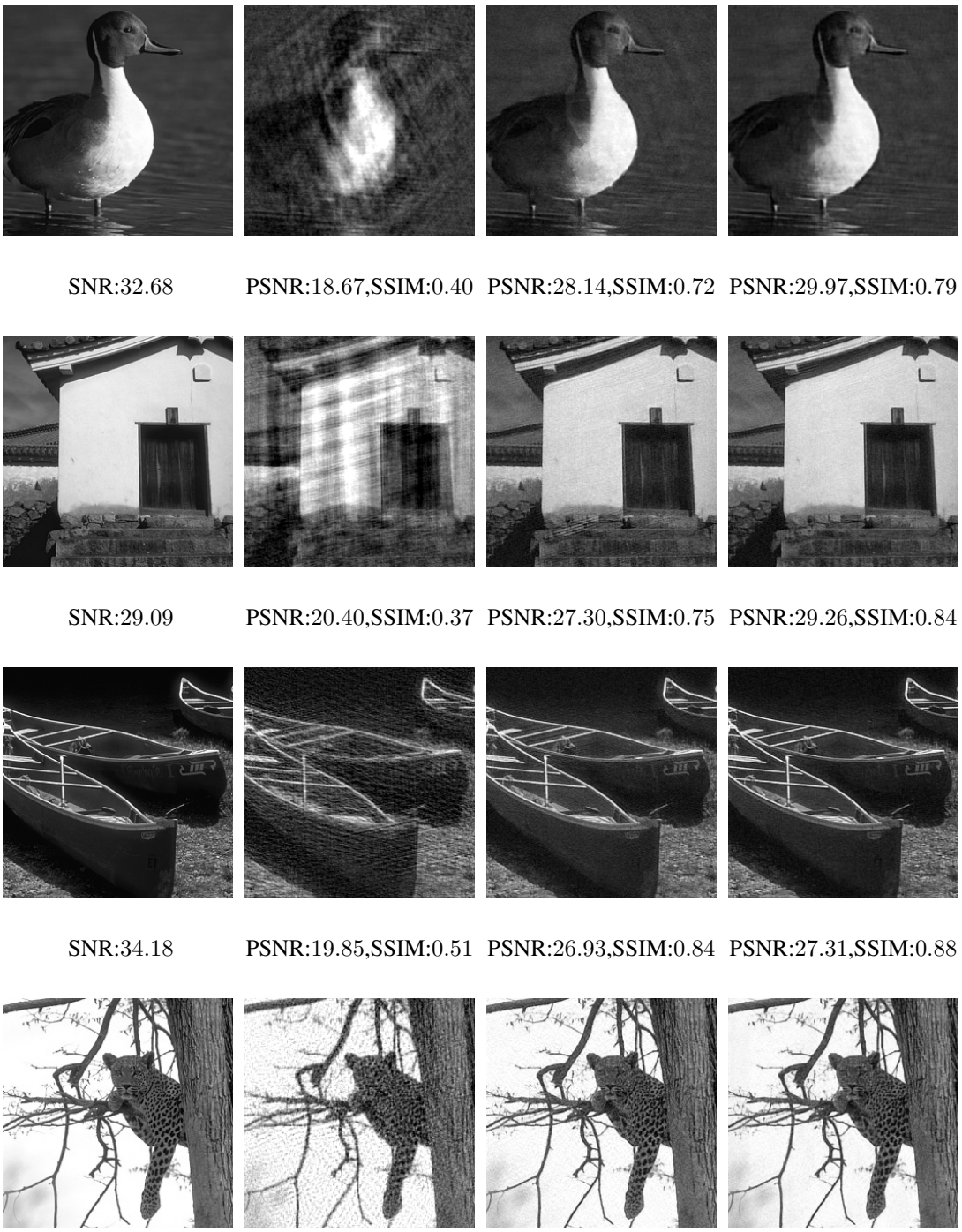


Figure B.1: The outputs of various algorithms for different test set images subjected to $\alpha = 3$ noise.

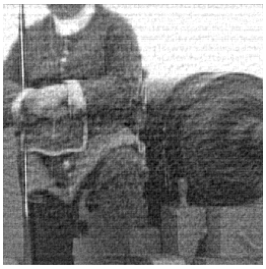
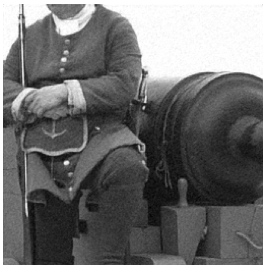
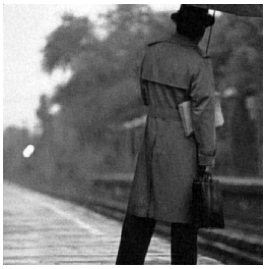
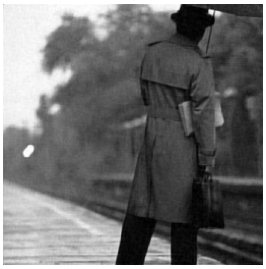
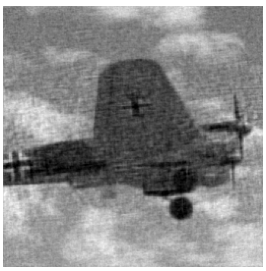
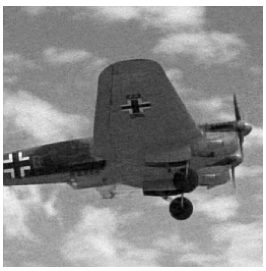
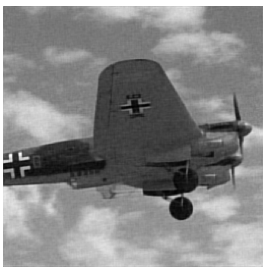
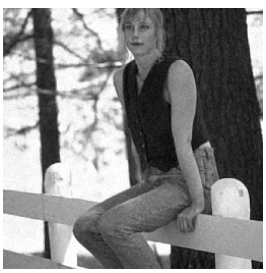
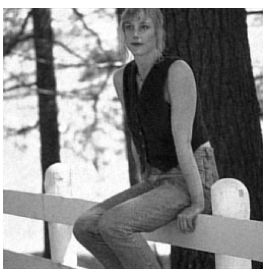
(a) Ground truth	(b) Initialization procedure	(c) InDI-PR ($T = 4, 2p = 1$)	(d) InDI-PR ($T = 32, 2p = 24$)
SNR:31.00	PSNR:24.11,SSIM:0.49	PSNR:31.15,SSIM:0.81	PSNR:32.28,SSIM:0.88
			
SNR:31.37	PSNR:26.21,SSIM:0.60	PSNR:32.60,SSIM:0.85	PSNR:34.21,SSIM:0.90
			
SNR:32.35	PSNR:26.95,SSIM:0.59	PSNR:32.53,SSIM:0.81	PSNR:34.80,SSIM:0.90
			
SNR:32.09	PSNR:25.67,SSIM:0.63	PSNR:32.25,SSIM:0.87	PSNR:33.83,SSIM:0.91
			

Figure B.2: The outputs of various algorithms for different test set images subjected to $\alpha = 3$ noise.

APPENDIX C

PROOFS FOR DDRM-PR

Definition C.1 (Original Form of DDRM)**.** *DDRM is a procedure utilizing a pre-trained unconditional diffusion model to solve any linear inverse problem in the following form:*

$$\mathbf{y} = \boldsymbol{H}\boldsymbol{x} + \mathbf{z} \quad \textit{where} \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{y}}^2 \mathbf{I}). \tag{C.1}$$

Using the singular value decomposition, i.e., $\boldsymbol{H} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathbf{T}}$ *with* $\mathbf{U}\mathbf{U}^{\mathbf{T}} = \mathbf{I}$ *and* $\mathbf{V}\mathbf{V}^{\mathbf{T}} = \mathbf{I}$*, we can make spectral definitions:*

$$\boldsymbol{H}^{\dagger} = \mathbf{V}\boldsymbol{\Sigma}^{\dagger}\mathbf{U}^{\mathbf{T}} \tag{C.2}$$

$$\bar{\mathbf{x}}_t = \mathbf{V}^{\mathbf{T}}\mathbf{x}_{\mathbf{t}} \tag{C.3}$$

$$\bar{\mathbf{y}} = \boldsymbol{\Sigma}^{\dagger}\mathbf{U}^{\mathbf{T}}\mathbf{y} \tag{C.4}$$

$$\bar{\mathbf{x}}_{\theta,t} = \mathbf{V}^{\mathbf{T}}\mathbf{x}_{\theta,t} \quad \textit{where} \quad \mathbf{x}_{\theta,t} = \mathbf{f}_{\theta}^{(t+1)}(\mathbf{x}_{t+1}) \textit{ is a denoiser with parameters } \theta. \tag{C.5}$$

Then, the DDRM sampling procedure is given by:

$$p_{\theta}^{(T)}\left(\overline{\mathbf{x}}_T^{(i)} \mid \mathbf{y}\right) = \begin{cases} \mathcal{N}\left(\overline{\mathbf{y}}^{(i)}, \sigma_T^2 - \frac{\sigma_{\mathbf{y}}^2}{s_i^2}\right) & \textit{if } s_i > 0 \\ \mathcal{N}\left(0, \sigma_T^2\right) & \textit{if } s_i = 0 \end{cases}$$

$$p_{\theta}^{(t)}\left(\overline{\mathbf{x}}_t^{(i)} \mid \mathbf{x}_{t+1}, \mathbf{y}\right) = \begin{cases} \mathcal{N}\left(\overline{\mathbf{x}}_{\theta,t}^{(i)} + \sqrt{1-\eta^2}\sigma_t \frac{\overline{\mathbf{x}}_{t+1}^{(i)} - \overline{\mathbf{x}}_{\theta,t}^{(i)}}{\sigma_{t+1}}, \eta^2\sigma_t^2\right) & \textit{if } s_i = 0 \\ \mathcal{N}\left(\overline{\mathbf{x}}_{\theta,t}^{(i)} + \sqrt{1-\eta^2}\sigma_t \frac{\overline{\mathbf{y}}^{(i)} - \overline{\mathbf{x}}_{\theta,t}^{(i)}}{\sigma_{\mathbf{y}}/s_i}, \eta^2\sigma_t^2\right) & \textit{if } \sigma_t < \frac{\sigma_{\mathbf{y}}}{s_i} \\ \mathcal{N}\left(\left(1-\eta_b\right)\overline{\mathbf{x}}_{\theta,t}^{(i)} + \eta_b\overline{\mathbf{y}}^{(i)}, \sigma_t^2 - \frac{\sigma_{\mathbf{y}}^2}{s_i^2}\eta_b^2\right) & \textit{if } \sigma_t \geq \frac{\sigma_{\mathbf{y}}}{s_i} \end{cases} \tag{C.6}$$

Here, we denote the i*-th index of any vector* $\mathbf{x}$ *by* $\mathbf{x}^{(i)}$*.*

Also, note that $\mathbf{x}_{\theta,t} = \mathbf{f}_{\theta}^{(t+1)}(\mathbf{x}_{t+1})$ *is trained with the regular unconditional diffusion*

process due to the conjugate variational distribution satisfying similar properties:

$$q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)=\mathcal{N}\left(\mathbf{x}_0, \sigma_t^2 \boldsymbol{I}\right) \quad \textit{with} \quad 0=\sigma_0<\sigma_1<\ldots .<\sigma_T \tag{C.7}$$

Definition C.2. *Let* $\alpha_t=\frac{1}{1+\sigma_t^2}$ *for all* t*. Equivalently,*

$$\sigma_t=\sqrt{\frac{1-\alpha_t}{\alpha_t}}, \quad \forall t. \tag{C.8}$$

Definition C.3. *Let* $\mathbf{x_t}=\sqrt{\alpha_t}\mathbf{x_t}$*, for all* t*. Cyan color will be used to denote this scaling.*

Lemma C.1.

$$\frac{\mathbf{x_t}-\sqrt{\alpha_t}\mathbf{x_0}}{\sqrt{1-\alpha_t}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \forall t. \tag{C.9}$$

Proof. From the assumed $q(\mathbf{x_t}, \mathbf{x_0})$, we know that

$$\frac{\mathbf{x_t}-\mathbf{x_0}}{\sigma_t} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{C.10}$$

Using Definitions C.2 and C.3,

$$\frac{\mathbf{x_t}-\mathbf{x_0}}{\sigma_t}=\frac{\frac{\mathbf{x_t}}{\sqrt{\alpha_t}}-\frac{\mathbf{x_0}}{\sqrt{\alpha_0}}}{\sqrt{\frac{1-\alpha_t}{\alpha_t}}}. \tag{C.11}$$

From Definition C.1, σ_0 is assumed to be 0, then, $\alpha_0=1$, and, it follows that

$$\frac{\mathbf{x_t}-\sqrt{\alpha_t}\mathbf{x_0}}{\sqrt{1-\alpha_t}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{C.12}$$

□

Corollary C.1.1. *If we have a perfect estimator of* $\mathbf{x}_0$ *denoted by* $\mathbf{x}_{\theta,t}=\mathbf{f}_\theta^{(t+1)}(\mathbf{x}_{t+1})$*, then,*

$$\epsilon_\theta^{(t+1)}(\mathbf{x}_{t+1})=\frac{\mathbf{x_{t+1}}-\sqrt{\alpha_{t+1}}\mathbf{x}_{\theta,t}}{\sqrt{1-\alpha_{t+1}}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{C.13}$$

Lemma C.2. *If* $\eta \sim \mathcal{N}(\mu, \mathbf{\Sigma})$*, then,* $\mathbf{A}\eta \sim \mathcal{N}(\mathbf{A}\mu, \mathbf{A\Sigma A^T})$.

Corollary C.2.1. *If* $\eta \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$*, then,* $\mathbf{V}\eta \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ *for an orthogonal matrix* $\mathbf{V}$.

Lemma C.3 (Reparametrization trick)**.** *If* $w \sim \mathcal{N}(\mu, \sigma^2)$*, then, we can write it as*

$$w=\mu+\sigma\epsilon \quad \textit{where} \quad \epsilon \sim \mathcal{N}(0,1). \tag{C.14}$$

Lemma C.4.

$$\boldsymbol{H}^\dagger \boldsymbol{H}=(\mathbf{V\Sigma^\dagger U^T})(\mathbf{U\Sigma V^T})=\mathbf{V\Sigma^\dagger \Sigma V^T}=\mathbf{\Sigma^\dagger \Sigma V V^T}=\mathbf{\Sigma^\dagger \Sigma} \tag{C.15}$$

Proof. Matrix multiplication with a square diagonal matrix is commutative. □

Theorem C.5 (Simplified form of DDRM)**.** *Under a noiseless setting, i.e., $\sigma_{\mathbf{y}}^2 = \mathbf{0}$, the overall DDRM process for linear inverse problems can be simplified to*

$$\begin{aligned}\mathbf{x}_t' &= \mathbf{x}_{\theta,t} - \boldsymbol{H}^\dagger \boldsymbol{H}\mathbf{x}_{\theta,t} + \boldsymbol{H}^\dagger \mathbf{y} \\ \mathbf{x}_t &= \sqrt{\alpha_t}\left(\eta_b \mathbf{x}_t' + (1-\eta_b)\,\mathbf{x}_{\theta,t}\right) + \sqrt{1-\alpha_t}\left(\eta\epsilon_t + (1-\eta)\epsilon_\theta^{(t+1)}\left(\mathbf{x}_{t+1}\right)\right)\end{aligned} \tag{C.16}$$

In this context, $\boldsymbol{H}^\dagger$ represents the Moore-Penrose pseudo-inverse of $\boldsymbol{H}$. The term $\mathbf{x}_{\theta,t} = f_\theta^{(t+1)}\left(\mathbf{x}_{t+1}\right)$ denotes the output of the denoising model at iteration $t+1$, while $\epsilon_\theta^{(t+1)}\left(\mathbf{x}_{t+1}\right) = \frac{\mathbf{x}_{t+1} - \sqrt{\alpha_{t+1}}\mathbf{x}_{\theta,t}}{\sqrt{1-\alpha_{t+1}}}$ indicates the estimated noise value. The constants η and η_b are hyperparameters defined by the user, and $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is a vector drawn from a standard Gaussian distribution [92].

Proof. We start with the original form of DDRM:

$$p_\theta^{(t)}\left(\overline{\mathbf{x}}_t^{(i)} \mid \mathbf{x}_{t+1}, \mathbf{y}\right) = \begin{cases} \mathcal{N}\left(\overline{\mathbf{x}}_{\theta,t}^{(i)} + \sqrt{1-\eta^2}\sigma_t \frac{\overline{\mathbf{x}}_{t+1}^{(i)} - \overline{\mathbf{x}}_{\theta,t}^{(i)}}{\sigma_{t+1}}, \eta^2\sigma_t^2\right) & \text{if } s_i = 0 \\ \mathcal{N}\left(\overline{\mathbf{x}}_{\theta,t}^{(i)} + \sqrt{1-\eta^2}\sigma_t \frac{\overline{\mathbf{y}}^{(i)} - \overline{\mathbf{x}}_{\theta,t}^{(i)}}{\sigma_{\mathbf{y}}/s_i}, \eta^2\sigma_t^2\right) & \text{if } \sigma_t < \frac{\sigma_{\mathbf{y}}}{s_i} \\ \mathcal{N}\left(\left(1-\eta_b\right)\overline{\mathbf{x}}_{\theta,t}^{(i)} + \eta_b\overline{\mathbf{y}}^{(i)}, \sigma_t^2 - \frac{\sigma_{\mathbf{y}}^2}{s_i^2}\eta_b^2\right) & \text{if } \sigma_t \geq \frac{\sigma_{\mathbf{y}}}{s_i} \end{cases} \tag{C.17}$$

Since $\sigma_{\mathbf{y}}^2 = \mathbf{0}$, the second case does not occur:

$$p_\theta^{(t)}\left(\overline{\mathbf{x}}_t^{(i)} \mid \mathbf{x}_{t+1}, \mathbf{y}\right) = \begin{cases} \mathcal{N}\left(\overline{\mathbf{x}}_{\theta,t}^{(i)} + \sqrt{1-\eta^2}\sigma_t \frac{\overline{\mathbf{x}}_{t+1}^{(i)} - \overline{\mathbf{x}}_{\theta,t}^{(i)}}{\sigma_{t+1}}, \eta^2\sigma_t^2\right) & \text{if } s_i = 0 \\ \mathcal{N}\left(\left(1-\eta_b\right)\overline{\mathbf{x}}_{\theta,t}^{(i)} + \eta_b\overline{\mathbf{y}}^{(i)}, \sigma_t^2\right) & \text{otherwise} \end{cases} \tag{C.18}$$

Use the reparametrization trick given in Lemma C.3:

$$\overline{\mathbf{x}}_t^{(i)} = \begin{cases} \overline{\mathbf{x}}_{\theta,t}^{(i)} + \sqrt{1-\eta^2}\sigma_t \frac{\overline{\mathbf{x}}_{t+1}^{(i)} - \overline{\mathbf{x}}_{\theta,t}^{(i)}}{\sigma_{t+1}} + \sqrt{\eta^2\sigma_t^2}\epsilon_t^{(i)} & \text{if } s_i = 0 \\ \left(1-\eta_b\right)\overline{\mathbf{x}}_{\theta,t}^{(i)} + \eta_b\overline{\mathbf{y}}^{(i)} + \sqrt{\sigma_t^2}\epsilon_t'^{(i)} & \text{otherwise} \end{cases} \tag{C.19}$$

where $\epsilon_t', \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Note that the matrix $\boldsymbol{\Sigma}^\dagger\boldsymbol{\Sigma}$ is a diagonal matrix with zeros at positions corresponding to zero singular values and ones elsewhere. This allows us to express $\overline{\mathbf{x}}_t$ in a more

compact form:

$$\begin{aligned}\overline{\mathbf{x}}_t = (\mathbf{I} - \boldsymbol{\Sigma}^\dagger\boldsymbol{\Sigma})\left(\overline{\mathbf{x}}_{\theta,t} + \sqrt{1-\eta^2}\sigma_t\frac{\overline{\mathbf{x}}_{t+1} - \overline{\mathbf{x}}_{\theta,t}}{\sigma_{t+1}} + \eta\sigma_t\epsilon_t\right)\\ +\boldsymbol{\Sigma}^\dagger\boldsymbol{\Sigma}\left((1-\eta_b)\,\overline{\mathbf{x}}_{\theta,t} + \eta_b\overline{\mathbf{y}} + \sigma_t\epsilon_t^{'}\right)\end{aligned} \tag{C.20}$$

Replace with the spectral definitions given in Definition C.1:

$$\begin{aligned}\mathbf{V}^\mathbf{T}\mathbf{x}_t = (\mathbf{I} - \boldsymbol{\Sigma}^\dagger\boldsymbol{\Sigma})\left(\mathbf{V}^\mathbf{T}\mathbf{x}_{\theta,t} + \sqrt{1-\eta^2}\sigma_t\frac{\mathbf{V}^\mathbf{T}\mathbf{x}_{t+1} - \mathbf{V}^\mathbf{T}\mathbf{x}_{\theta,t}}{\sigma_{t+1}} + \eta\sigma_t\epsilon_t\right)\\ +\boldsymbol{\Sigma}^\dagger\boldsymbol{\Sigma}\left((1-\eta_b)\,\mathbf{V}^\mathbf{T}\mathbf{x}_{\theta,t} + \eta_b\boldsymbol{\Sigma}^\dagger\mathbf{U}^\mathbf{T}\mathbf{y} + \sigma_t\epsilon_t^{'}\right)\end{aligned} \tag{C.21}$$

Multiply both sides by $\mathbf{V}$. And, recall that multiplication with square diagonal matrices is commutative:

$$\begin{aligned}\mathbf{x}_t = (\mathbf{I} - \boldsymbol{\Sigma}^\dagger\boldsymbol{\Sigma})\left(\mathbf{x}_{\theta,t} + \sqrt{1-\eta^2}\sigma_t\frac{\mathbf{x}_{t+1} - \mathbf{x}_{\theta,t}}{\sigma_{t+1}} + \eta\sigma_t\mathbf{V}\epsilon_t\right)\\ +\boldsymbol{\Sigma}^\dagger\boldsymbol{\Sigma}\left((1-\eta_b)\,\mathbf{x}_{\theta,t} + \eta_b\boldsymbol{H}^\dagger\mathbf{y} + \sigma_t\mathbf{V}\epsilon_t^{'}\right)\end{aligned} \tag{C.22}$$

Use Lemma C.4 ($\boldsymbol{H}^\dagger\boldsymbol{H} = \boldsymbol{\Sigma}^\dagger\boldsymbol{\Sigma}$) and Corollary C.2.1 ($\mathbf{V}\epsilon_t = \epsilon_t$):

$$\begin{aligned}\mathbf{x}_t = (\mathbf{I} - \boldsymbol{H}^\dagger\boldsymbol{H})\left(\mathbf{x}_{\theta,t} + \sqrt{1-\eta^2}\sigma_t\frac{\mathbf{x}_{t+1} - \mathbf{x}_{\theta,t}}{\sigma_{t+1}} + \eta\sigma_t\epsilon_t\right)\\ +\boldsymbol{H}^\dagger\boldsymbol{H}\left((1-\eta_b)\,\mathbf{x}_{\theta,t} + \eta_b\boldsymbol{H}^\dagger\mathbf{y} + \sigma_t\epsilon_t^{'}\right)\end{aligned} \tag{C.23}$$

Use Definitions C.3 and C.2:

$$\begin{aligned}\frac{\mathbf{x}_t}{\sqrt{\alpha_t}} = (\mathbf{I} - \boldsymbol{H}^\dagger\boldsymbol{H})\left(\mathbf{x}_{\theta,t} + \sqrt{1-\eta^2}\sqrt{\frac{1-\alpha_t}{\alpha_t}}\frac{\frac{\mathbf{x}_{t+1}}{\sqrt{\alpha_{t+1}}} - \mathbf{x}_{\theta,t}}{\sqrt{\frac{1-\alpha_{t+1}}{\alpha_{t+1}}}} + \eta\sqrt{\frac{1-\alpha_t}{\alpha_t}}\epsilon_t\right)\\ +\boldsymbol{H}^\dagger\boldsymbol{H}\left((1-\eta_b)\,\mathbf{x}_{\theta,t} + \eta_b\boldsymbol{H}^\dagger\mathbf{y} + \sqrt{\frac{1-\alpha_t}{\alpha_t}}\epsilon_t^{'}\right)\end{aligned} \tag{C.24}$$

Simplify:

$$\begin{aligned}\mathbf{x}_t = (\mathbf{I} - \boldsymbol{H}^\dagger\boldsymbol{H})\left(\sqrt{\alpha_t}\mathbf{x}_{\theta,t} + \sqrt{1-\eta^2}\sqrt{1-\alpha_t}\frac{\mathbf{x}_{t+1} - \sqrt{\alpha_{t+1}}\mathbf{x}_{\theta,t}}{\sqrt{1-\alpha_{t+1}}} + \eta\sqrt{1-\alpha_t}\epsilon_t\right)\\ +\boldsymbol{H}^\dagger\boldsymbol{H}\left(\sqrt{\alpha_t}\,(1-\eta_b)\,\mathbf{x}_{\theta,t} + \sqrt{\alpha_t}\eta_b\boldsymbol{H}^\dagger\mathbf{y} + \sqrt{1-\alpha_t}\epsilon_t^{'}\right)\end{aligned} \tag{C.25}$$

Rearrange the terms:

$$\begin{aligned}\mathbf{x}_t =&\sqrt{\alpha_t}\mathbf{x}_{\theta,t} + \sqrt{\alpha_t}(-1+(1-\eta_b))\boldsymbol{H}^\dagger\boldsymbol{H}\mathbf{x}_{\theta,t} + \sqrt{\alpha_t}\eta_b\boldsymbol{H}^\dagger\mathbf{y}\\ &+\eta\sqrt{1-\alpha_t}\epsilon_t\\ &+\boldsymbol{H}^\dagger\boldsymbol{H}(-\eta+1)\sqrt{1-\alpha_t}\epsilon_t'\\ &+(\mathbf{I}-\boldsymbol{H}^\dagger\boldsymbol{H})\sqrt{1-\eta^2}\sqrt{1-\alpha_t}\frac{\mathbf{x}_{t+1}-\sqrt{\alpha_{t+1}}\mathbf{x}_{\theta,t}}{\sqrt{1-\alpha_{t+1}}}\end{aligned} \tag{C.26}$$

Use Corollary C.1.1 and approximate $\sqrt{1-\eta^2} \approx 1-\eta$ for $\eta \in [0,1]$.

$$\begin{aligned}\mathbf{x}_t =&\sqrt{\alpha_t}\left(\mathbf{x}_{\theta,t} + \eta_b\left(\mathbf{x}_{\theta,t} - \boldsymbol{H}^\dagger\boldsymbol{H}\mathbf{x}_{\theta,t} + \boldsymbol{H}^\dagger\mathbf{y}\right) - \eta_b\mathbf{x}_{\theta,t}\right)\\ &+\sqrt{1-\alpha_t}\eta\epsilon_t\\ &+\boldsymbol{H}^\dagger\boldsymbol{H}(1-\eta)\sqrt{1-\alpha_t}\epsilon_\theta^{(t+1)}\left(\mathbf{x}_{t+1}\right)\\ &+(\mathbf{I}-\boldsymbol{H}^\dagger\boldsymbol{H})(1-\eta)\sqrt{1-\alpha_t}\epsilon_\theta^{(t+1)}\left(\mathbf{x}_{t+1}\right)\\ =&\sqrt{\alpha_t}\left(\eta_b\left(\mathbf{x}_{\theta,t} - \boldsymbol{H}^\dagger\boldsymbol{H}\mathbf{x}_{\theta,t} + \boldsymbol{H}^\dagger\mathbf{y}\right) + (1-\eta_b)\mathbf{x}_{\theta,t}\right)\\ &+\sqrt{1-\alpha_t}\left(\eta\epsilon_t + (1-\eta)\epsilon_\theta^{(t+1)}\left(\mathbf{x}_{t+1}\right)\right)\end{aligned} \tag{C.27}$$

□