Unrolled Wirtinger Flow With Deep Decoding Priors for Phaseless Imaging

Samia Kazemi, Bariscan Yonel, *Member, IEEE*, and Birsen Yazici, *Fellow, IEEE*

***Abstract*—We introduce a deep learning (DL) based network and an associated exact recovery theory for imaging from intensity-only measurements. The network architecture uses a recurrent structure that unrolls the Wirtinger Flow (WF) algorithm with a deep decoding prior that enables performing the algorithm updates in a lower dimensional encoded image space. We use a separate deep network (DN), referred to as the encoding network, for transforming the spectral initialization used in the WF algorithm to an appropriate initial value for the encoded domain. The unrolling scheme models a fixed number of iterations of the underlying optimization algorithm into a recurrent neural network (RNN). Furthermore, it facilitates simultaneous learning of the parameters of the decoding and encoding networks and the RNN. We establish a sufficient condition to guarantee exact recovery under deterministic forward models. Additionally, we demonstrate the relation between the Lipschitz constants of the trained decoding prior and encoding networks to the convergence rate of the WF algorithm. We show the practical applicability of our method in synthetic aperture imaging using high fidelity simulation data from the PCSWAT software. Our numerical study shows that the decoding prior and the encoding network facilitate improvements in sample complexity.**

***Index Terms*—Deep learning, inverse problems, phase retrieval, deep prior, Wirtinger Flow, synthetic aperture imaging, algorithm unrolling.**

I. Introduction

A. Motivation and Prior Art

Phaseless imaging refers to the task of reconstructing an image from measurements whose magnitude or intensity values are available while the phase information is either missing or unreliable. This challenging problem necessitates compensation either through hand-crafted prior information [1] or significant measurement redundancy [2], [3]. In practical imaging applications with deterministic forward maps, hand-crafted priors may not be sufficiently descriptive of the underlying image domain to reduce the requirement of a large number of measurements [4], [5]. In this paper, we introduce a deep learning (DL) based phaseless imaging method that incorporates data-driven prior information for deterministic imaging problems with theoretical convergence and an exact recovery guarantee.

We consider the state-of-the-art phase retrieval methods that fall under two general categories: *Wirtinger Flow* (WF) type algorithms [2], [6]–[8] and *DL-based approaches* [9]–[13]. The first category includes WF [2] and its variants [4]–[7], [14] which offer exact recovery guarantees based on non-convex optimization. Unlike the earlier lifting-based convex phase-retrieval algorithms [15], [16], WF performs iterations in the signal space relieving the extensive computation and memory requirements. However, classical WF requires an appropriate choice of the initial estimate, learning rate and high sample complexity of $\mathcal{O}(N \log N)$ under the Gaussian measurement model. Several initial estimates for WF have been studied including the spectral estimation [2], spectral estimation with sample truncation [17] and more general sample processing functions [18]–[20], linear spectral estimation [21], orthogonality-promoting initialization [22] etc. Original WF algorithm has been extended to include prior information [4], [7] to reduce its sample complexity, most prominent of which is sparsity. However, finding a hand-crafted optimal basis over which the unknown image is sparse can be challenging. Other variants of WF aiming to reduce sample complexity include [6], [14]. However, the exact recovery theory of WF [2] and its variants [4], [6], [7], [14] relies on the assumption that the forward map is Gaussian. This poses a fundamental limitation for imaging applications since the forward models are almost always deterministic.

Recently, in [8], we introduced a mathematical framework for establishing an exact recovery guarantee for the WF algorithm involving deterministic forward maps under a sufficient condition that sets a concentration bound on the spectral matrix [2]. This paves the way for the adoption of WF-type algorithms in a wide range of practical applications with provable performance guarantees. However, this framework does not account for prior information about the image domain or study how the sufficient condition will be affected by the incorporation of such information.

The second category of state-of-the-art methods for phaseless imaging are practically attractive as they present a trade-off between the number of measurements and the training data, by solving the imaging problem in a lower dimensional encoded image space using a generative prior [9]–[12]. These are iterative algorithms where the parameters of the prior network, often

Manuscript received 10 August 2021; revised 6 March 2022 and 5 June 2022; accepted 22 June 2022. Date of publication 8 July 2022; date of current version 10 August 2022. This work was supported in part by the Air Force Office of Scientific Research (AFOSR) under Grant Agreement FA9550-19-1-0284, in part by Office of Naval Research (ONR) under Grant Agreement N00014-18-1-2068, in part by the National Science Foundation (NSF) under Grant ECCS-1809234, and in part by the United States Naval Research Laboratory (NRL) under Grant Agreement N00173-21-1-G007. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ivan W. Selesnick. *(Corresponding author: Birsen Yazici.)*

The authors are with the Department of Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180 USA (e-mail: kazems@rpi.edu; yonelb2@rpi.edu; yazici@ecse.rpi.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TCI.2022.3189217, provided by the authors.

Digital Object Identifier 10.1109/TCI.2022.3189217

referred to as the generative network, are learned to capture the global characteristics of the image manifold. Once trained, starting from a randomly initialized encoded image, this network is used to update the encoded image estimation. A convergence guarantee for the phaseless imaging problem is established for real positive-valued unknown image components in [9] given that the trained weight matrices and the forward map satisfy a weight distribution condition and a range restricted concentration property, respectively. For the same network as in [9] with expansive layers of particular dimensionalities, and a measurement matrix and trained weight matrices of i.i.d. Gaussian distributed components, [11] shows that optimal sample complexity can be achieved for the phaseless imaging problem after a sufficient number of iterations. However, since the prior network is trained separately from the phaseless imaging problem, these methods require large training sets in order to effectively estimate the probability distribution over the image domain instead of a conditional distribution given the phaseless measurements [23]. Additionally, this training scheme precludes the inclusion of an optimal initialization scheme for the encoded image space.

For overcoming the large training set requirement and fixed image space restriction of the generative prior, a related class of methods utilizes untrained networks in which the network structure itself works as the prior [24]. For the phaseless imaging problem, a deep decoder [13], [25], which uses an under-parameterized architecture, is utilized in [13] and an exact recovery guarantee is established for a two-layer decoder model and Gaussian distributed forward map that satisfies a specific restricted eigenvalue condition. However, an optimal initialization scheme for the weights of the network, instead of the encoded image, is not established. Additionally, theoretical results for this approach are very limited.

To address the limitations of state-of-the-art phaseless imaging methods, in this paper, we combine the WF algorithm and theory in [8] with a DL-based approach. We consider the following two major modifications: the use of a deep decoding prior in conjunction with DL-based initialization and the unrolling of the WF algorithm into a recurrent neural network (RNN) architecture which enables end-to-end training. Our overall network is composed of the transformation network for initialization referred to as the *encoder*, an RNN that represents the unrolled gradient descent updates of the WF in the encoded domain and the deep decoding prior network referred to as the *decoder*.

Unrolling, which has been widely implemented to a range of linear inversion problems [23], [26] has limited utilization in the phase retrieval literature. In [27], an unrolled network is introduced for a Fourier phase retrieval problem with a reference signal. In [28], a complex unrolled network with unsupervised training is proposed for lensless microscopy imaging from phaseless measurements. An unrolled Incremental Reshaped Wirtinger Flow based phase retrieval approach is presented in [29] for direct image estimation from amplitude measurements. However, the trainable parameter set for this method is only related to the learning rates and no theoretical exact recovery guarantee is established. To the best of our knowledge, our approach is the first to unroll a phaseless imaging algorithm with deep priors and end-to-end supervised training for general imaging applications. Additionally, we have established a theoretical exact recovery guarantee. A related approach in [30] incorporates adaptive step sizes, but their implementation does not use a fixed number of iterations, the step sizes are not learned and no theoretical exact recovery guarantee is established.

B. Our Approach and Its Advantages

Our approach bridges the class of theoretically sound state-of-the-art purely optimization-based non-convex approaches with data-driven schemes deploying deep decoding priors for phaseless imaging in a deterministic setting. Instead of the generative adversarial network (GAN) [31] based training used in the prior work [9]–[12], we adopt an end-to-end training approach where the parameters of the decoder, RNN and the encoder are learned simultaneously during training. The unrolling strategy benefits from the inherent computational efficiency of a trained optimal network. Additionally, being derived from model-based iterative algorithms, the network also offers interpretability of its architecture and parameters unlike an arbitrary deep network for phaseless imaging.

Our approach relates the spectral initialization-based WF algorithm with a generative prior based approach within a DL framework. Existing applications of the generative prior [9]–[12] lack a rigorous justification for the choice of initialization. Furthermore, it is not well-understood how this value affects the convergence rate. By establishing an explicit connection to the spectral initialization step, we determine the effect of the decoding network on the validity of the convergence guarantees and the rate of convergence to the true solution. Our theoretical analysis reveals two key observations:

- Firstly, the parameters of the underlying encoding and decoding prior networks have direct implications on the convergence rate and initialization accuracy which can be quantified by their Lipschitz constant values after training. A learned decoding prior can achieve a faster convergence rate compared to non-DL based WF [8] as long as certain Lipschitz constant related conditions are satisfied by the trained networks.
- Secondly, using the lower dimensional embedding of the decoding prior, we establish a new sufficient condition for exact recovery where, by virtue of specific imposed conditions on the decoder, the concentration property considered in [8] is parameterized over the encoded space. Hence, a sufficiently accurate initial estimate for the algorithm can be obtained using fewer measurements, as the representations are embedded in the lower dimensional space by the encoder. This sample complexity reduction aspect is also observed empirically through our numerical simulations.

The main differences with the existing generative prior based phase retrieval methods are notably in the initialization criteria, and the type of conditions assumed on the measurement vectors and the DL network parameters for establishing exact recovery guarantee when compared to [9], [11]. In [9]–[12], the encoded unknown is randomly initialized, while in our approach, which can be viewed as a DL enhanced WF, we implement a DL network to transform the spectral initialization output to an encoded

TABLE I
List of Important Notations

$\hat{\boldsymbol{\rho}}$	Estimated image
$\boldsymbol{\rho}^*$	True unknown image
$\boldsymbol{\rho}\boldsymbol{\rho}^H$	Lifted image vector $\boldsymbol{\rho}$
$\mathbf{d}$	Vector of the measured intensity values $\{d_m\}_{m=1}^M$
$\mathcal{F}$	Lifted forward map where $\mathbf{d} = \mathcal{F}(\boldsymbol{\rho}^* \boldsymbol{\rho}^{*H})$
$\mathbf{F}$	Forward map with $\{\mathbf{a}_m^H\}_{m=1}^M$ along its rows
$\mathcal{F}^H$	ℓ_2 adjoint of $\mathcal{F}$
$\mathbf{Y}$	Spectral matrix defined as $\mathbf{Y} := \frac{1}{M}\mathcal{F}^H(\mathbf{d})$
$\mathcal{G}$	Encoding network or Encoder
$\mathcal{H}$	Decoding prior network or Decoding network or Decoder
$\mathbf{E}_\rho$	$\mathcal{H}(\mathbf{y})\mathcal{H}(\mathbf{y})^H - \boldsymbol{\rho}^*\boldsymbol{\rho}^{*H}$
$\langle.,.\rangle_F$	Frobenius inner product

initialization value in order to facilitate a better starting point. Even though the spectral initialization is computationally more expensive compared to a random initialization step, imaging applications in [10], [12] use multiple initial guesses each of which is iteratively updated for selecting the best one. Our approach avoids the need for repeating the algorithm for an arbitrary number of initial guesses, and its computation complexity is of the same order as in [32]. Additionally, unlike [9], [11], our sufficient conditions on the trained DL networks for achieving exact recovery guarantee do not depend on the explicit consideration of the network architectures or imposition of specific properties on the trained network weights. The sufficient condition on the forward map is similar to the deterministic WF analysis in [8] rather than the generative network architecture dependent condition in [9], [11].

Our numerical simulation results demonstrate the ability of end-to-end learning with the unrolled WF method for reconstructing a wide range of unknown image sets. This includes MNIST image set of handwritten digits, simulated images with geometric objects and PCSWAT [33] simulated images with mine-like objects for different non-Gaussian deterministic forward maps.

C. Notation and Organization of the Paper

Bold upper case and bold lower case letters are used to represent matrices and vectors, respectively. $\|\mathbf{X}\|_F$ refers to the Frobenius norm of $\mathbf{X}$, and it is calculated as Tr$(\mathbf{X}^H\mathbf{X})$. Tr(.) denotes the trace of a matrix, while superscripts T and H on a matrix (or vector) denote its transpose and Hermitian transpose, respectively. $\|.\|$ around a matrix and a vector refer to their spectral norm and ℓ_2-norm, respectively. Calligraphic letters and doublestruck upper case letters are used for operators and sets, respectively. We use lower case Greek letters to represent various constants, and lower case italic letters, with or without subscripts, are used to denote different functions. For a network $\mathcal{B}$ with input $\mathbf{x}$, $\mathcal{B}(\mathbf{x})$ is its output vector. Finally, we are using upper case italic letters for constant integers, and a set of integer values from 1 to K is written as $[K]$.

Table I includes a list of important notations used throughout this paper and the supplementary material.

The rest of the paper is organized as follows: The problem statement and background on the non-DL based phase retrieval methods are discussed in Section II. The DL-based overall imaging network is introduced in Section III. Theoretical foundations required for establishing the exact recovery guarantee of our approach are discussed in Section IV. Section V presents our theoretical results involving the accuracy of the DL-based initial value, convergence guarantee and properties on the DNs for desired reconstruction performance. The training process and the implementation details of specific properties of the encoder, decoder and the RNN are presented in Subsection VI-A and Subsection VI-B discusses the computational complexity of our approach. Numerical simulations examining the performance of our approach compared to the WF algorithm and other DL-based methods are presented in Section VII. Finally, Section VIII concludes the paper.

II. Problem Statement

A. The Phase Retrieval Problem

The phase retrieval problem entails estimating an unknown $\boldsymbol{\rho}^* \in \mathbb{C}^N$, from its intensity, or magnitude-only measurements of the form:

$$d_m = |\langle \mathbf{a}_m, \boldsymbol{\rho}^* \rangle|^2, \quad \text{for } m = 1, 2, \ldots M, \tag{1}$$

where $\mathbf{a}_m \in \mathbb{C}^N$, for all $m = 1, \ldots, M$, denotes the m^{th} sampling vector. These vectors constitute a known, *linear* measurement model, $\mathbf{F}$, pertaining to the application of interest, such as Gaussian sampling, coded diffraction patterns, Fourier transform etc. We refer to $\mathbf{F}$ as the forward map. When $\{\mathbf{a}_m\}_{m=1}^M$ are Fourier sampling vectors, the problem is classically known as Fourier phase retrieval, or *the phase problem* in optical imaging, and quantum physics fields.

Fundamentally, (1) constitutes a system of M quadratic equations, and solving it is known to be NP-hard in general [34]. Nonetheless, classical algorithms based on alternating minimization have been used to empirical success in optical imaging applications [35]–[37], despite the severe ill-posedness of the problem that arises due to the quadratic dependence of the measurements to the quantity of interest in (1) [38].

Over the last decade, optimization-based approaches have methodically progressed towards establishing performance guarantees in exactly recovering $\boldsymbol{\rho}^*$ from $\mathbf{d} = [d_1, \cdots d_M]^T \in \mathbb{R}^M$. First major developments to this end have been through a reformulation of (1) via *lifting* the problem, as the recovery of a rank-1, positive semidefinite (PSD) unknown $\boldsymbol{\rho}^*\boldsymbol{\rho}^{*H}$ from $\mathbf{d}$. (1) become equivalent to realizations under a *linear* measurement model, governed by a *lifted forward map*, $\mathcal{F} : \mathbb{C}^{N\times N} \mapsto \mathbb{C}^M$, where

$$d_m = \langle \mathbf{a}_m \mathbf{a}_m^H, \boldsymbol{\rho}^*\boldsymbol{\rho}^{*H} \rangle_F, \ \text{ for } m = 1, \ldots M. \tag{2}$$

This reformulation facilitates the use of established tools from low rank matrix recovery theory through convex-relaxations and semidefinite programming [15], [16]. The injectivity and the spectral properties of $\mathcal{F}$ over rank-1, PSD matrices therefore determine the exact recovery of $\boldsymbol{\rho}^*\boldsymbol{\rho}^{*H}$ [15].

More recently, algorithms that attain performance guarantees by directly operating on the original signal domain [39]–[41] have been introduced to overcome the demanding computational and memory requirements of the lifting-based approaches. One of the most prominent one is the WF algorithm [2], which minimizes the following functional:

$$\mathcal{J}(\boldsymbol{\rho}) := \frac{1}{2\,M} \sum_{m=1}^{M} |(\mathbf{a}_m)^H \boldsymbol{\rho}\boldsymbol{\rho}^H \mathbf{a}_m - d_m|^2. \tag{3}$$

At the p^{th} iteration step, the WF algorithm updates the current estimate $\boldsymbol{\rho}^{(p-1)}$ of the unknown quantity as follows:

$$\boldsymbol{\rho}^{(p)} = \boldsymbol{\rho}^{(p-1)} - \frac{\gamma_p}{\|\boldsymbol{\rho}^{(0)}\|^2} \nabla \mathcal{J}(\boldsymbol{\rho})_{\boldsymbol{\rho}=\boldsymbol{\rho}^{(p-1)}}. \tag{4}$$

Here, γ_p denotes the learning rate at the p^{th} stage and the gradient is given by the Wirtinger derivative of $\mathcal{J}(\boldsymbol{\rho})$,

$$\nabla \mathcal{J}(\boldsymbol{\rho}) = \left(\frac{\partial \mathcal{J}}{\partial \boldsymbol{\rho}}\right)^H. \tag{5}$$

The critical component of the WF framework is at the initialization step, where $\boldsymbol{\rho}^{(0)}$ is determined from the leading eigenvector $\mathbf{v}_0$ of the backprojection estimate $\mathbf{Y} \in \mathbb{C}^{N \times N}$ as follows:

$$\mathbf{Y} := \frac{1}{M} \mathcal{F}^H(\mathbf{d}), \tag{6}$$

$$\boldsymbol{\rho}^{(0)} = \sqrt{\lambda_0}\mathbf{v}_0, \tag{7}$$

where $\mathcal{F}^H$ is the ℓ_2 adjoint of $\mathcal{F}$ and the scaling factor $\sqrt{\lambda_0}$ is a norm-estimate of the unknown image of interest. We refer to $\mathbf{Y}$ as the spectral matrix.

Under the following concentration inequality on $\mathbf{Y}$

$$\|\mathbf{Y} - \left(\boldsymbol{\rho}\boldsymbol{\rho}^H + \|\boldsymbol{\rho}\|^2 \mathbf{I}\right)\| \leq \delta \|\boldsymbol{\rho}\|^2, \tag{8}$$

the initial estimate provably enters a basin of attraction in the neighborhood of the global solution set $\mathbb{P} := \{\boldsymbol{\rho}^* e^{i\phi}, \phi \in [0, 2\pi)\}$, such that convergence is guaranteed under the validity of a *regularity condition* for the loss functional $\mathcal{J}$ in the noise-free setting with Gaussian sampling, and coded-diffraction models [16]. These amount to exact recovery guarantees in the statistical setting, where any $\boldsymbol{\rho} \in \mathbb{C}^N$ can be exactly recovered up to a global phase factor, with overwhelming probability if the number of samples exceeds $\mathcal{O}(N \log N)$.

On the other hand in [8], the validity of (8) for all $\boldsymbol{\rho} \in \mathbb{C}^N$ with a sufficiently small δ (< 0.184) was shown to be a sufficient condition for universal exact recovery via WF for any $\mathcal{F}$ in a deterministic mathematical framework. Hence, deterministic forward maps, $\mathbf{F}$, that relate to underlying data collection geometry are equipped with exact recovery guarantees. This is especially useful for wave-based imaging applications, where the sampling vectors, $\{\mathbf{a}_m\}_{m=1}^M$, are related to the transmitter and receiver locations, transmission signal waveform, and its speed within the propagation medium, and are unlikely to follow i.i.d. Gaussian distribution.

B. WF With a Deep Decoding Prior

In this paper, we build on the mathematical arguments introduced in [8] in establishing the exact recovery guarantee for a DL-based algorithm. This allows our DL-based algorithm and theoretical results to be applicable to a wide range of practical imaging applications involving deterministic forward maps. In particular, we present our phaseless imaging approach that performs WF iterations in *a lower dimensional encoded space* in $\mathbb{C}^{N_y}$, where $N_y \ll N$, in lieu of the original image domain in $\mathbb{C}^N$.

The key distinction from existing phase retrieval theory arises from the *non-linearity* of the underlying measurement map prior to loss of phase information, since (1) corresponds to $\mathbf{d} = |\mathbf{F}\boldsymbol{\rho}^*|^2$, where $|\cdot|$ denotes element-wise absolute-value operation, and $\mathbf{F} \in \mathbb{C}^{M \times N}$ is the matrix with $\{\mathbf{a}_m^H\}_{m=1}^M$ as its rows. Namely, we now assume that our image class of interest resides in a low dimensional manifold $\mathbb{T}$, embedded in the high dimensional space in $\mathbb{C}^N$. We aim to capture this image manifold $\mathbb{Y}$ by parameterization over the $\mathbb{C}^{N_y}$ in the range of a non-linear transformation $\mathcal{H} : \mathbb{Y} \subset \mathbb{C}^{N_y} \mapsto \mathbb{T}$, which we refer to as the *decoder*. This yields a measurement model of the form:

$$d_m = |\langle \mathbf{a}_m, \mathcal{H}(\mathbf{y}^*) \rangle|^2, \quad \text{for } m = 1, \ldots M \tag{9}$$

where $\boldsymbol{\rho}^* = \mathcal{H}(\mathbf{y}^*)$, such that we have a compositely non-linear mapping, $\mathbf{d} = |\mathbf{F}\mathcal{H}(\mathbf{y}^*)|^2$, over the low dimensional parameter space in $\mathbb{Y} \subset \mathbb{C}^{N_y}$.

The problem consists of two key elements: *i*) given $\mathcal{H}$, solving for the underlying, *compressive* representation $\mathbf{y} \in \mathbb{Y}$ from (9), and *ii*) solving for an $\mathcal{H}$ that sufficiently approximates the image manifold $\mathbb{T} \subset \mathbb{C}^N$. While the first component requires the composite mapping formed by $\mathbf{F}$ and $\mathcal{H}$ to demonstrate favorable properties of the parameter space, the other requires constructing one such representation in the first place. Practically, the two can be summarized under an objective using a training set of $\mathbb{D} := \{\boldsymbol{\rho}_t^*, \mathbf{d}_t\}_{t=1}^T$, such that

$$\begin{aligned} \underset{\{\mathbf{y}_t\}_{t=1}^T, \mathcal{H} \in \mathbb{W}}{\arg\min} \; & \frac{1}{TM} \sum_{t=1}^{T} \sum_{m=1}^{M} |\mathbf{a}_m^H \mathcal{H}(\mathbf{y}_t) \mathcal{H}(\mathbf{y}_t)^H \mathbf{a}_m - d_{t,m}|^2 \\ \text{s.t. } & \|\mathcal{H}(\mathbf{y}_t) - \boldsymbol{\rho}_t^*\| \leq \varepsilon, \; \forall t = 1, \ldots T, \end{aligned} \tag{10}$$

where $\mathbb{W}$ denotes a space of functionals that acts as a constraint in the search of $\mathcal{H}$, and $\varepsilon > 0$ models the approximation error in the range of the decoder.

Ultimately, despite serving as a conceptual motivation, solving (10) is not meaningful without attaining proper generalization over the image manifold $\mathbb{T}$, i.e., any $\boldsymbol{\rho} \in \mathbb{T}$ must be reliably reconstructed by recovering its encoded representation from its intensity-only measurements. To this end, we enlist a DL-based approach, where $\mathcal{H}$ is obtained in a *task-driven* manner, such that it facilitates the accurate recovery of elements in $\mathbb{T}$ in its range after the iterative procedure of WF is deployed on the lower dimensional, encoded parameter space. The DL-based approach effectively *splits* the objective in (10) to be minimized over its forward, and back-propagation stages. Namely, at the forward pass, we pursue a solution $\hat{\mathbf{y}} \in \mathbb{Y}$ that minimizes the following objective function for each training sample:

$$\mathcal{K}(\mathbf{y}) := \frac{1}{2\,M} \sum_{m=1}^{M} \left[(\mathbf{a}_m)^H \mathcal{H}(\mathbf{y}) \mathcal{H}(\mathbf{y})^H \mathbf{a}_m - d_m\right]^2, \tag{11}$$

whereas in the back-propagation, we use the solution $\hat{\mathbf{y}}$ to formulate the training loss over $\mathcal{H} \in \mathbb{W}$, evaluated over the training set $\mathbb{D}$.

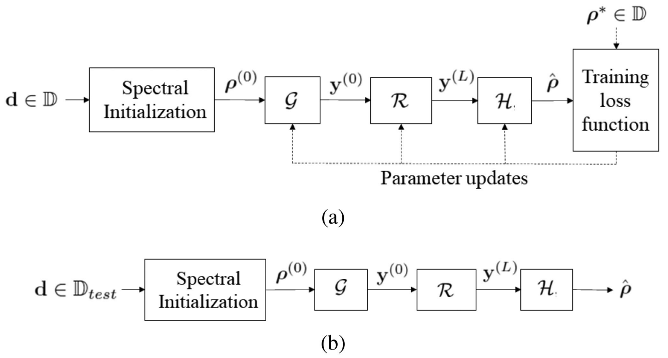
Accordingly, our approach incorporates *a deep decoding prior* into the WF framework. Deep decoding prior refers to the type of compressive representation implemented under our decoding network $\mathcal{H}$, as it constrains the reconstructed images to its output space. We opt to use decoding prior in referring to $\mathcal{H}$ to differentiate our overall approach from works that consider generative priors [9]–[12], which do not utilize the phaseless measurements and the corresponding ground truth images for training, and instead use pre-trained GAN generator model for $\mathcal{H}$. The end-to-end training of $\mathcal{H}$ transforms an M/N phase retrieval problem into an M/N_y phase retrieval problem akin to the generative prior setting. Hence, the composite operator mapping $\mathbf{y}$ to the measurements attains a higher oversampling factor, albeit, at the cost of non-linearity. On the other hand, overcoming the N/N_y factor reduction is offloaded to the approximation capability of the decoder. In accordance, we are interested in the theoretical justifications of recovering a true representation $\mathbf{y}^* \in \mathbb{Y}$, for a given $\mathcal{H}$ such that $\mathcal{H}(\mathbf{y}^*) = \boldsymbol{\rho}^*$, using the iterative scheme of WF.

Unlike [9], [11], our architecture is based on the observation that $\mathcal{H}$ and the measurement map $\mathcal{F}$ need to satisfy certain sufficient conditions for exact recovery *in composition with each other*. This serves as our key motivation to utilize end-to-end training, as it directly entangles the presence of the generator with the measurement map of the problem, hence drives the training procedure to enhance the feasibility of the phase retrieval problem over $\mathbb{T}$. However, guarantees on finding such an $\mathcal{H}$, or the impact of approximation and generalization errors encountered in the training of $\mathcal{H}$ are beyond the scope of this paper.

III. Network Architecture

As our phaseless imaging approach recovers an encoded version of the unknown image through WF updates, our first challenge is to design an efficient initialization scheme for the encoded image space. To this end, we utilize the spectral initialization step, and learn a non-linear transformation from the set of initial estimates $\mathbb{S} \subseteq \mathbb{C}^N$ to the set of encoded initial estimates $\mathbb{Y}_0 \subseteq \mathbb{C}^{N_y}$, $\mathcal{G} : \mathbb{S} \mapsto \mathbb{Y}_0$, to map the spectral estimate $\boldsymbol{\rho}^{(0)} \in \mathbb{S}$ to an initial estimate $\mathbf{y}^{(0)} \in \mathbb{Y}_0$ in the encoded image space. We refer to $\mathcal{G}$ as the *encoding network*. We use an L-layer RNN, $\mathcal{R}$, to generate the final estimated encoded image $\mathcal{R}(\mathbf{y}^{(0)}) = \mathbf{y}^{(L)} = \hat{\mathbf{y}}$, where $\mathbf{y}^{(L)}$ is the output of the L^{th} layer of the RNN. We denote the set of encoded image values generated at the l^{th} RNN layer by $\mathbb{Y}_l \subset \mathbb{C}^{N_y}$ for $l \in [L-1]$ and define $\mathbb{Y}$ as $\mathbb{Y} = \bigcup_{l=0}^{L} \mathbb{Y}_l \subset \mathbb{C}^{N_y}$. Thus, $\mathcal{R} : \mathbb{Y}_0 \mapsto \mathbb{Y}$. Finally, the output from the RNN is decoded back by $\mathcal{H} : \mathbb{Y} \mapsto \mathbb{T}$ to generate the estimated image $\hat{\boldsymbol{\rho}} \in \mathbb{T}$. Under exact recovery, $\hat{\boldsymbol{\rho}} = \boldsymbol{\rho}^*$.

In our network architecture, the encoder, RNN and the decoder are jointly learned through supervised training. The training dataset $\mathbb{D}$ is composed of different ground truth or correct images and the corresponding intensity measurement vectors.



(a)

(b)

Fig. 1. Schematic diagrams showing the (a) training and (b) inversion processes.

On the other hand, each new sample from the test set, $\mathbb{D}_{test}$, only requires the intensity measurement vector which is then applied to the trained imaging network to produce the estimated image. A block diagram of the training and inversion phases of our algorithm are shown in Fig. 1(a) and 1(b), respectively.

A. RNN Structure From the Iterative WF Updates

Starting from the initial encoded representation $\mathbf{y}^{(0)}$, iterative WF update at the l^{th} stage is calculated as follows:

$$\mathbf{y}^{(l)} = \mathbf{y}^{(l-1)} - \frac{\gamma_l}{\|\mathbf{y}^{(0)}\|^2} \nabla\mathcal{K}(\mathbf{y})_{\mathbf{y}=\mathbf{y}^{(l-1)}}. \tag{12}$$

$\mathbf{y}^{(l)}$ denotes the output at the l^{th} iteration and γ_l is a positive real-valued constant associated with the learning rate for the l^{th} update. The WF update in (12) results in $\mathbf{y}^{(l)}$ that reduces the data fidelity term $\mathcal{K}(.)$ compared to $\mathbf{y}^{(l-1)}$. The gradient of $\mathcal{K}(\mathbf{y})$ with respect to $\mathbf{y} \in \mathbb{C}^{N_y}$ is given by

$$\nabla\mathcal{K}(\mathbf{y}) = \left(\frac{\partial\mathcal{K}}{\partial\mathbf{y}}\right)^H = \frac{1}{M}\nabla\mathcal{H}(\mathbf{y})\mathcal{F}^H(\mathbf{e})\,\mathcal{H}(\mathbf{y}), \tag{13}$$

where $\mathbf{e} = [e_1 \quad \cdots \quad e_M]$ and $e_m \in \mathbb{R}$, for $m \in [M]$, is defined as $e_m := \mathbf{a}_m^H \mathcal{H}(\mathbf{y})\mathcal{H}(\mathbf{y})^H \mathbf{a}_m - d_m$.

Instead of continuing to update the encoded representation until convergence, we consider a fixed number of iterative update steps over which the algorithm is promoted to recover accurate solutions over certain conditions on the network parameters. Similar to [42]–[46], L number of subsequent update steps from (12) are mapped into the stages of an L-layer RNN. The resulting network is referred to as an RNN due to the recursive nature of its architecture. Each RNN layer essentially carries out a WF update on the encoded representation. The learning rate related constants, $\{\gamma_l\}_{l=1}^{L}$, are all trainable parameters of the RNN whose values are learned during the training process. The overall diagram of our DL-based inversion network for phaseless imaging is shown in Fig. 2.

B. Lipschitz Constants of the DL Networks

The encoding and decoding prior networks are trained with the goal of recovering images from their low dimensional representations at a faster convergence rate compared to the WF

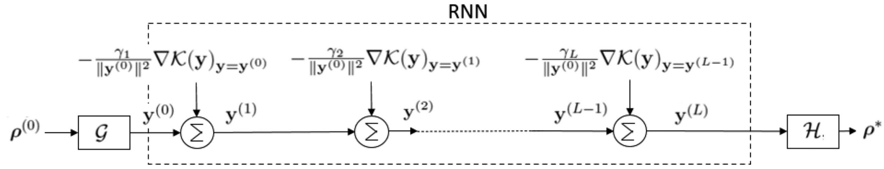

RNN
$-\frac{\gamma_1}{\|\mathbf{y}^{(0)}\|^2}\nabla\mathcal{K}(\mathbf{y})_{\mathbf{y}=\mathbf{y}^{(0)}}$
$-\frac{\gamma_2}{\|\mathbf{y}^{(0)}\|^2}\nabla\mathcal{K}(\mathbf{y})_{\mathbf{y}=\mathbf{y}^{(1)}}$
$-\frac{\gamma_L}{\|\mathbf{y}^{(0)}\|^2}\nabla\mathcal{K}(\mathbf{y})_{\mathbf{y}=\mathbf{y}^{(L-1)}}$
$\boldsymbol{\rho}^{(0)}$
$\mathcal{G}$
$\mathbf{y}^{(0)}$
Σ
$\mathbf{y}^{(1)}$
Σ
$\mathbf{y}^{(2)}$
$\mathbf{y}^{(L-1)}$
Σ
$\mathbf{y}^{(L)}$
$\mathcal{H}$
$\boldsymbol{\rho}^{*}$

Fig. 2. Block diagram of DL-based phaseless imaging network.

algorithm and extending the recovery guarantees of [8] to challenging problem settings with $M<N$ for arbitrary forward maps. In order to achieve the above two objectives, we characterize the impact of the encoder and decoder networks on recovery guarantees through their Lipschitz constants, rather than the explicit architectures of the networks or any probabilistic properties on their learned parameter values. Appropriate ranges of these constants that are associated with improved recovery performance compared to the WF algorithm is presented in Section V.

The Lipschitz constant of $\mathcal{G}$ is defined as the smallest value of $\mu_{\mathcal{G}} \in \mathbb{R}^+$ satisfying [47]

$$\frac{\|\mathcal{G}(\boldsymbol{\rho}_1^{(0)}) - \mathcal{G}(\boldsymbol{\rho}_2^{(0)})\|}{\|\boldsymbol{\rho}_1^{(0)} - \boldsymbol{\rho}_2^{(0)}\|} \leq \mu_{\mathcal{G}}, \tag{14}$$

$\forall \boldsymbol{\rho}_1^{(0)}, \boldsymbol{\rho}_2^{(0)} \in \mathbb{S}$, and is given by

$$\mu_{\mathcal{G}} = \sup_{\boldsymbol{\rho}^{(0)}\in\mathbb{S}} \sigma(\nabla\mathcal{G}(\boldsymbol{\rho}^{(0)})), \tag{15}$$

where $\sigma(\mathbf{A})$ denotes the largest singular value of $\mathbf{A}$.

Suppose $\mathbf{y}^{(0)} = \mathcal{G}(\boldsymbol{\rho}^{(0)})$ is expressed as a function of a set of weight matrices $\mathbf{U}_j \in \mathbb{C}^{P_j \times P_{j-1}}$'s, bias vectors $\mathbf{b}_j \in \mathbb{C}^{P_j}$'s, and non-linear functions $f_j(.)$'s, where $j \in [J]$, with $P_0 = N$ and $P_J = N_y$. The output at the j^{th} step, denoted by $\breve{\mathbf{y}}_j \in \mathbb{C}^{P_j}$, relates to its input $\breve{\mathbf{y}}_{j-1} \in \mathbb{C}^{P_{j-1}}$ as follows:

$$\breve{\mathbf{y}}_j = f_j(\mathbf{U}_j\breve{\mathbf{y}}_{j-1} + \mathbf{b}_j), \tag{16}$$

where $\breve{\mathbf{y}}_J = \mathbf{y}^{(0)}$ and $\breve{\mathbf{y}}_0 = \boldsymbol{\rho}^{(0)}$. The activation function $f_j(.)$ operates componentwise on the corresponding vector inputs. For the choice of $f_j(.)$ as the rectified linear unit (ReLU), Lipschitz constant of $f_j(.)$ is upper-bounded by 1 and thus, the Lipschitz constant of $\mathcal{G}(.)$ for this case is upper bounded by $\prod_{j=1}^{J} \sigma(\mathbf{U}_j)$. Similarly, the Lipschitz constant of $\mathcal{H}$ is calculated as $\mu_{\mathcal{H}} = \sup_{\mathbf{y}\in\mathbb{Y}} \sigma(\nabla\mathcal{H}(\mathbf{y}))$.

IV. Theoretical Foundations

In order to justify the effectiveness of our phaseless imaging approach, we provide a theoretical foundation towards attaining exact recovery for a given, arbitrary lifted forward map $\mathcal{F}$, and an image manifold that is assumed to be characterized in the range of a non-linear operator $\mathcal{H}$.

In terms of the technical content of the exact recovery theory, our work differs from prior works in [9], [11] in two notable ways. The first pertains to the conditions exerted on $\mathcal{H}$. In [9], [11], a pre-determined architecture is assumed for $\mathcal{H}$ and a concentration property on the network weights is used to facilitate recovery guarantees by a sufficient condition on $\mathcal{F}$. We do not deploy an architecture specification for $\mathcal{H}$, and only assume a *local* concentration-type property instead. The second pertains to the sufficient condition on the measurement map $\mathcal{F}$, where [9], [11] use a *range restricted* RIP-type property on the underlying linear sampling vectors $\mathbf{F}$, while our sufficient condition enforces a range restriction on the sufficient condition introduced in [8]. The major distinction arises in the domain of the accompanying concentration property, where our work evades the requirement of validity over pair-wise differences.

A. Approach

To understand the feasibility of such a theoretical justification under an arbitrary pairing of $\mathcal{F}$ and $\mathcal{H}$, it is useful to initially revisit the standard phase retrieval problem in the statistical setting of Gaussian sampling. Indeed, theoretical results in phase retrieval literature commonly consider this case, where $\mathbf{a}_m$ are i.i.d. complex Gaussian distributed, with which the recovery from intensity-only measurements is achieved with overwhelming probability [2], [7].

Using the property that Gaussian distribution is invariant under unitary transformations, the classically studied statistical phase retrieval problem under the Gaussian sampling model is equivalent to a $1D$-Fourier phase retrieval problem under a *linear Gaussian generator:*

$$\mathbf{d} = |\mathbf{A}\mathbf{s}|^2 = |\mathbf{F}_M\mathbf{F}_M^H\mathbf{A}\mathbf{s}|^2 = |\mathbf{F}_M\tilde{\mathbf{A}}\mathbf{s}|^2 = |\mathbf{F}_M\mathbf{t}|^2, \tag{17}$$

where $\mathbf{A} \in \mathbb{C}^{M\times N_s}$ has all i.i.d. Gaussian distributed components, $\mathbf{s} \in \mathbb{C}^{N_s}$, $\mathbf{F}_M \in \mathbb{C}^{M\times M}$ is the discrete 1D-Fourier matrix, $\tilde{\mathbf{A}} = \mathbf{F}_M^H\mathbf{A}$ and $\mathbf{t} = \tilde{\mathbf{A}}\mathbf{s}$. In other words, standard statistical theory states that a signal $\mathbf{t} \in \mathbb{C}^M$ realized from a Gaussian generative prior can provably be recovered from its $M-$point periodogram, *if the intrinsic dimension N_s is sufficiently low*.

Exact recovery guarantees in the statistical setting highlight the power of having *a generative prior* at inference, albeit disguised as the measurement model due to spherical symmetry of the Gaussian distribution. This is because the $1D$-Fourier phase retrieval problem is well-known to be severely ill-posed: it admits at best 2^M non-equivalent solutions in the feasible set of $\mathbf{d} = |\mathbf{F}_M\mathbf{t}|^2$ for an arbitrary $\mathbf{t} \in \mathbb{C}^M$ [48]. The linear Gaussian generator alleviates the fundamental limitations in this regard, and provides a guarantee directly on the lower, N_s-dimensional encoded space, given that $\mathbf{t} = \tilde{\mathbf{A}}\mathbf{s}$.

Ultimately, our work aims at generalizing this phenomenon by: *i*) using the deterministic setting of [8] to account for an arbitrary $\mathcal{F}$, and *ii*) incorporating the presence of a non-linear $\mathcal{H}$ that can capture the signal domain. To this end, we quantify the impact of operating in the lower dimensional encoded domain on the existing deterministic guarantees of [8] by specifying conditions on $\mathcal{H}$ within the sufficient conditions, and identifying the numerical impact of the generator, i.e. our decoder, on convergence guarantees.

B. Background

a) Exact phase retrieval theory: For universality of exact recovery described in [8], the concentration bound in (8) is a sufficient condition if it holds over all $\boldsymbol{\rho} \in \mathbb{C}^N$ with $\delta < 0.184$. The terms involved in the concentration bound are relevant for the initial estimate to land within a basin of attraction around the true solution $\boldsymbol{\rho}^*$, guaranteeing that:

$$\frac{1}{M}\|\mathcal{F}(\boldsymbol{\rho}\boldsymbol{\rho}^H - \boldsymbol{\rho}^*\boldsymbol{\rho}^{*H})\|^2 \geq (1-\delta_1^{WF})\|\boldsymbol{\rho}\boldsymbol{\rho}^H - \boldsymbol{\rho}^*\boldsymbol{\rho}^{*H}\|_F^2. \tag{18}$$

Let $\mathbb{N}_\epsilon(\boldsymbol{\rho}^*)$ denote this ϵ-neighborhood of $\boldsymbol{\rho}^*$ obtained from the sufficient condition in (8), and $\epsilon \in \mathbb{R}^+$ and $\delta_1^{WF} \in \mathbb{R}^+$ are both functions of δ. In the end, (18) facilitates the restricted strong convexity around the solution $\boldsymbol{\rho}^*$ if $\delta_1^{WF} < 1$, which (8) guarantees an initial estimate to land in for any $\boldsymbol{\rho}^*$. The way to establish (8) as a sufficient condition is through deriving (18) as a deterministic consequence, and showing that the requirement of $\delta_1^{WF} < 1$ implies $\delta < 0.184$ in the sufficient condition.

b) Range restriction with $\mathcal{H}$: The stringency of the sufficient condition in (8) arises through its universality over all $\boldsymbol{\rho} \in \mathbb{C}^N$ and the corresponding requirement for $\mathcal{F}^H\mathcal{F}$ to be well-conditioned over the manifold of rank-1 PSD matrices. On the other hand, with the presence of $\mathcal{H}$, the range of the decoder incorporates an additional constraint, and hence, creates a smaller feasible set for the problem over which $\mathcal{F}^H\mathcal{F}$ should be well-conditioned instead.

An intuitive incorporation of the image manifold in the recovery guarantees therefore is by restricting the parameter space of the original concentration bound, where the lifted normal operator is to satisfy, for all $\mathbf{y} \in \mathbb{Y} \subset \mathbb{C}^{N_y}$:

$$\left\|\frac{1}{M}\mathcal{F}^H\mathcal{F}(\mathcal{H}(\mathbf{y})\mathcal{H}(\mathbf{y})^H) - \left(\mathcal{H}(\mathbf{y})\mathcal{H}(\mathbf{y})^H + \|\mathcal{H}(\mathbf{y})\|^2\mathbf{I}\right)\right\| \\ \leq \delta\|\mathcal{H}(\mathbf{y})\|^2. \tag{19}$$

(19) shows that the concentration property of $\mathcal{F}^H\mathcal{F}$ is now required to hold over only the image manifold captured by the range of $\mathcal{H}$.

To fully understand the usefulness of this condition, we must establish its corresponding restricted strong convexity property over the image manifold. Namely, for a $\boldsymbol{\rho} = \mathcal{H}(\mathbf{y})$, and a ground truth $\boldsymbol{\rho}^* = \mathcal{H}(\mathbf{y}^*)$, does (19) with a sufficiently small δ imply the property in (18) with δ_1^{WF} replaced by $\delta_1 < 1$ in some locality in the *parameter space*, i.e. $\mathbf{y} \in \mathbb{N}_{\epsilon_\mathbf{y}}(\mathbf{y}^*)$? Here, $\mathbb{N}_{\epsilon_\mathbf{y}}(\mathbf{y}^*)$ denotes the $\epsilon_\mathbf{y}$-neighborhood of $\mathbf{y}^*$ and $\epsilon_\mathbf{y} \in \mathbb{R}^+$.

c) The limitation for sufficiency: In order to verify whether the restricted concentration property is sufficient, we consider first the linear perturbation operator Δ that maps $\boldsymbol{\rho}\boldsymbol{\rho}^H$ to $\mathbb{C}^{N\times N}$ over all $\boldsymbol{\rho}$ vectors that are reproducible by the decoder from $\mathbf{y} \in \mathbb{Y}$, as

$$\Delta(\boldsymbol{\rho}\boldsymbol{\rho}^H) = \frac{1}{M}\mathcal{F}^H\mathcal{F}(\boldsymbol{\rho}\boldsymbol{\rho}^H) - (\boldsymbol{\rho}\boldsymbol{\rho}^H + \|\boldsymbol{\rho}\|^2\mathbf{I}). \tag{20}$$

Similarly to the steps of the proof of Lemma III.4 in [8], it is easy to verify that the validity of the restricted strong convexity condition through (18) hinges on the concentration property of a perturbation operator Δ, over the pairwise differences,

$$\left|\langle\Delta(\mathbf{E}_{\boldsymbol{\rho}}), \mathbf{E}_{\boldsymbol{\rho}}\rangle_F\right| \leq \delta_1\,\|\mathbf{E}_{\boldsymbol{\rho}}\|_F^2, \tag{21}$$

where $\mathbf{E}_{\boldsymbol{\rho}}$ is defined as

$$\mathbf{E}_{\boldsymbol{\rho}} = \mathcal{H}(\mathbf{y})\mathcal{H}(\mathbf{y})^H - \boldsymbol{\rho}^*\boldsymbol{\rho}^{*H}, \tag{22}$$

and that (21) is guaranteed to hold with $\delta_1 < 1$ when (19) is satisfied. As we know, $|\langle\Delta(\mathbf{E}_{\boldsymbol{\rho}}), \mathbf{E}_{\boldsymbol{\rho}}\rangle_F|$ can be upper bounded by $\sqrt{2}\|\mathbf{E}_{\boldsymbol{\rho}}\|_F\|\Delta(\mathbf{E}_{\boldsymbol{\rho}})\|$. Moreover, $\|\Delta(\mathbf{E}_{\boldsymbol{\rho}})\|$ can be upper bounded by $\sum_{i=1}^{2}|\lambda_i|\|\Delta(\mathbf{v}_i\mathbf{v}_i^H)\|$, where $\lambda_i \in \mathbb{R}$ and $\mathbf{v}_i \in \mathbb{C}^N$ are the i^{th} eigenvalue and the corresponding eigenvector of $\mathbf{E}_{\boldsymbol{\rho}}$, respectively, for $i \in \{1, 2\}$.

Consequently, to promote (19) as a sufficient condition for our approach, $\mathbf{v}_i$'s need to be reproducible by the decoding network $\mathcal{H}$, such that $\|\Delta(\mathbf{v}_i\mathbf{v}_i^H)\|$ terms are controlled. For an arbitrary pair of $\boldsymbol{\rho}, \boldsymbol{\rho}^*$, the error $\mathbf{E}_{\boldsymbol{\rho}}$ for the corresponding lifted Kronecker signals admit a direct spectral analysis, such that the $\mathbf{v}_i$ are formed by *affine combinations* in the range of $\mathcal{H}$ (see Section IX-A of the supplementary material). This presents the key limitation for the sufficiency of a range restriction by the generator $\mathcal{H}$, unless the domain of concentration is expanded to include the union of pair-wise affine hulls of the elements in the range of $\mathcal{H}$.

C. Conditioning $\mathcal{H}$

a) Sufficiency with linearity: It is clear that for a linear $\mathcal{H}$, (19) is a sufficient condition, as the affine combinations are reproducible by $\mathcal{H}$ via an affine combination in the $\mathbb{Y}$-domain. However, for a general non-linear $\mathcal{H}$, the eigenvectors $\mathbf{v}_i$ do not necessarily admit such a representation. We instead are interested in casting (19) as a sufficient condition through specific conditions on an arbitrary, non-linear $\mathcal{H}$. To this end, we first identify the properties that facilitate our objective when using a linear decoder model, towards obtaining an intuitive extension onto the general case. The assumption that $\mathcal{H}$ is a linear map, i.e., $\mathcal{H}(\mathbf{y}) = \mathbf{H}\mathbf{y}$ where $\mathbf{H} \in \mathbb{C}^{N\times N_y}$, leads to

$$\|\Delta(\mathbf{E}_{\boldsymbol{\rho}})\| = \|\Delta(\mathbf{H}(\mathbf{y}\mathbf{y}^H - \mathbf{y}^*\mathbf{y}^{*H})\mathbf{H}^H)\|. \tag{23}$$

Now, $\mathbf{y}\mathbf{y}^H - \mathbf{y}^*\mathbf{y}^{*H}$ can be represented by its eigenvalues and eigenvectors as $\sum_{i=1}^{2}\lambda_i\mathbf{u}_i\mathbf{u}_i^H$, where $\lambda_i \in \mathbb{R}$ and $\mathbf{u}_i \in \mathbb{C}^{N_y}$ are eigenvalues and the corresponding eigenvectors for $i = 1, 2$. $\mathbf{u}_1$ and $\mathbf{u}_2$ are constructed from affine combinations of $\mathbf{y}$, $\mathbf{y}^*$ per spectral analysis presented in Section IX-A of the supplementary material.

b) Requirements for the general case: We now assume that (19) holds for all $\mathcal{H}(\mathbf{y})$, $\mathbf{y} \in \mathbb{R}^{N_y}$ for convenience. Therefore, since $\|\Delta(\mathbf{E}_{\boldsymbol{\rho}})\|$ can be upper bounded by $\sum_{i=1}^{2}|\lambda_i|\|\Delta((\mathbf{H}\mathbf{u}_i)(\mathbf{H}\mathbf{u}_i)^H)\|$ when $\mathcal{H}$ is linear, then using the relation in (19), we have from (23),

$$\|\Delta(\mathbf{E}_{\boldsymbol{\rho}})\| \leq \delta\sum_{i=1}^{2}|\lambda_i|\|\mathbf{H}\mathbf{u}_i\|^2. \tag{24}$$

Here, the first crucial property of $\mathcal{H}$ arises, as the Lipschitz continuity of $\mathcal{H}$, along with the assumption that $\mathcal{H}(\mathbf{0}) = \mathbf{0}$, which yields the following upper bound for linear $\mathcal{H}$:

$$\begin{aligned}\|\Delta(\mathbf{E}_{\boldsymbol{\rho}})\| &\le \delta \max(|\lambda_1|, |\lambda_2|) \sum_{i=1}^{2} \|\mathbf{H}\mathbf{u}_i\|^2 \\ &\le 2\delta \mu_{\mathcal{H}}^2 \|\mathbf{y}\mathbf{y}^H - \mathbf{y}^*\mathbf{y}^{*H}\|. \end{aligned} \quad (25)$$

Although this bound is not the tightest, it is of interest because, it gives a blueprint that befits generalization to the non-linear setting. Mainly, in the linear setting with a *spectrally well-conditioned* generator, we can obtain a universal constant (2 in this case) that upper bounds this perturbation operator only through the leading eigenvalue-eigenvector pair, since $\|\mathbf{H}\mathbf{u}_i\|^2 \le \mu_{\mathcal{H}}^2$ by the Lipschitz property of $\mathbf{H}$. The key observation is that via an encoder-decoder scheme that enforces the model to operate in an $\epsilon_{\mathbf{y}}$-neighborhood in the parameter space, such a condition as in (25) is only needed to be satisfied *locally* over $\mathbb{Y}$, in lieu of the global property demonstrated by a linear $\mathcal{H}$.

c) Extension via a local property: For a general non-linear decoder, we instead perform this analysis using an operator $\tilde{\mathcal{H}} : \mathbb{C}^{N_y \times N_y} \mapsto \mathbb{C}^{N \times N}$, which is defined as follows:

1) Given input $\mathbf{Z} \in \mathbb{C}^{N_y \times N_y}$, extract the leading eigenvalue-eigenvector pair: λ_0, $\mathbf{u}_0$.
2) Apply $\mathcal{H}$ on $\sqrt{\lambda_0}\mathbf{u}_0$ to calculate $\mathcal{H}(\sqrt{\lambda_0}\mathbf{u}_0)$.
3) Get output $\tilde{\mathcal{H}}(\mathbf{Z})$ by lifting: $\mathcal{H}(\sqrt{\lambda_0}\mathbf{u}_0)\mathcal{H}(\sqrt{\lambda_0}\mathbf{u}_0)^H$.

Under this definition, our desired bound on the perturbation operator for a generic $\mathcal{H}$ as can be written as

$$\|\Delta(\tilde{\mathcal{H}}(\mathbf{y}\mathbf{y}^H) - \tilde{\mathcal{H}}(\mathbf{y}^*\mathbf{y}^{*H}))\| \le \hat{\delta}\|\mathbf{y}\mathbf{y}^H - \mathbf{y}^*\mathbf{y}^{*H}\|_F, \quad (26)$$

which, incorporating the locality property on the encoded domain, should hold $\forall \mathbf{y}^* \in \mathbb{Y}$, $\mathbf{y} \in \mathbb{N}_{\epsilon_{\mathbf{y}}}(\mathbf{y}^*)$. For the PSD rank-1 inputs $\mathbf{y}\mathbf{y}^H$ and $\mathbf{y}^*\mathbf{y}^{*H}$, $\tilde{\mathcal{H}}(\mathbf{y}\mathbf{y}^H)$ and $\tilde{\mathcal{H}}(\mathbf{y}^*\mathbf{y}^{*H})$ are equal to $\mathcal{H}(\mathbf{y})\mathcal{H}(\mathbf{y})^H$ and $\boldsymbol{\rho}^*\boldsymbol{\rho}^{*H}$, respectively. Moreover, we are not necessarily interested in this bound globally as obtained for the linear case in (25), but only locally, since that is sufficient for our guarantees.

This leads us to the following property on $\mathcal{H}$: for the definition of $\tilde{\mathcal{H}}$ presented above, for a given $\mathcal{F}$, the following inequality is satisfied $\forall \mathbf{y}^* \in \mathbb{Y}$, $\mathbf{y} \in \mathbb{N}_{\epsilon_{\mathbf{y}}}(\mathbf{y}^*)$:

$$\|\Delta(\tilde{\mathcal{H}}(\mathbf{y}\mathbf{y}^H) - \tilde{\mathcal{H}}(\mathbf{y}^*\mathbf{y}^{*H}))\| \le \omega(\epsilon_{\mathbf{y}})\|\Delta(\tilde{\mathcal{H}}(\mathbf{y}\mathbf{y}^H - \mathbf{y}^*\mathbf{y}^{*H}))\|, \quad (27)$$

where $\omega(\epsilon_{\mathbf{y}})$ is a positive real-valued constant. We omit the term in the bracket for future references to this constant, and its dependency on $\epsilon_{\mathbf{y}}$ should be understood. Under this condition, it is straightforward to verify that the desired bound in (26) is satisfied with a constant

$$\hat{\delta} = \omega\mu_{\mathcal{H}}^2\delta, \quad (28)$$

as shown in Section IX-B of the supplementary file.

V. Recovery Guarantees

In this section, we present the exact recovery guarantee for our end-to-end DL-based algorithm. This result is built upon the theoretical foundations presented in Section IV. We elaborate on the numerical implications of our result, and discuss its key outcomes in quantifying the impact and limitations of incorporating a decoding prior. Finally, we consider the practical implications of our result for implementation purposes.

A. Main Result

Let $\mathrm{dist}(\mathbf{y}^{(0)}, \mathbf{y}^*)$ be the distance between $\mathbf{y}^{(0)}$ and $\mathbf{y}^*$ defined as follows:

$$\mathrm{dist}(\mathbf{y}^{(0)}, \mathbf{y}^*) = \min_{\phi \in [0,\pi]} \|\mathbf{y}^{(0)} - \mathbf{y}^* \mathrm{e}^{i\phi}\|. \quad (29)$$

Our main result concerns the convergence of the WF iterates to the true representation in the encoded space via our unrolled, encoder-decoder network architecture. Let $\mu_{\mathcal{G}}$, $\mu_{\mathcal{R}}$ and $\mu_{\mathcal{H}}$ be the Lipschitz constants of $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{H}$, respectively. We assume that there exists $\tilde{\mu}_{\mathcal{H}} > 0$ and $\mu_{\mathcal{H}} > 0$ such that

$$\tilde{\mu}_{\mathcal{H}} \le \frac{\|\mathcal{H}(\mathbf{y}_1) - \mathcal{H}(\mathbf{y}_2)\|}{\|\mathbf{y}_1 - \mathbf{y}_2\|} \le \mu_{\mathcal{H}}, \quad (30)$$

for all $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{Y}$.

We define $\epsilon_{\mathbf{y}}$, which we introduced in Subsection IV-B, as $\epsilon_{\mathbf{y}} := \chi\mu_{\mathcal{H}}\epsilon$. χ is a positive real-valued constant and ϵ is defined in (21) in [8]. χ is lower bounded by

$$\chi \ge \max\left[b_1(\mu_{\mathcal{G}}, \mu_{\mathcal{H}}, \epsilon), b_2(\mu_{\mathcal{G}}, \mu_{\mathcal{H}}, \mu_{\mathcal{R}}, \epsilon)\right], \quad (31)$$

and $b_1(\mu_{\mathcal{G}}, \mu_{\mathcal{H}}, \epsilon)$ and $b_2(\mu_{\mathcal{G}}, \mu_{\mathcal{H}}, \mu_{\mathcal{R}}, \epsilon)$ are defined in Section IX-C of the supplementary material along with the detailed derivation of (31). We also define the following quantities:

$$c(\delta, \epsilon_{\mathbf{y}}) := (1 + \epsilon_{\mathbf{y}})(2 + \epsilon_{\mathbf{y}})(2 + \omega\delta), \quad (32)$$

$$\epsilon_{\boldsymbol{\rho}} := \mu_{\mathcal{G}}\mu_{\mathcal{R}}\mu_{\mathcal{H}}(1 + \epsilon)\epsilon_{\mathbf{y}}, \quad (33)$$

$$\delta_1 := \frac{\sqrt{2}\hat{\delta}(2 + \epsilon_{\boldsymbol{\rho}})(2 + \epsilon_{\mathbf{y}})}{\tilde{\mu}_{\mathcal{H}}^2(1 - \epsilon_{\boldsymbol{\rho}})(2 - \epsilon_{\boldsymbol{\rho}})}, \quad (34)$$

$$h(\delta, \epsilon_{\mathbf{y}}) := \tilde{\mu}_{\mathcal{H}}^4(1 - \delta_1)(1 - \epsilon_{\boldsymbol{\rho}})(2 - \epsilon_{\boldsymbol{\rho}}). \quad (35)$$

Theorem 1: Suppose the conditions in (19) and (27) are satisfied for all $\mathbf{y} \in \mathbb{Y}$, where $\mathbb{Y}$ is an affine subset of $\mathbb{C}^{N_y}$. Additionally, assume that there exist $\tilde{\mu}_{\mathcal{H}} > 0$ and $\mu_{\mathcal{H}}$ such that (30) holds; and $\mathcal{G}(\mathbf{0}) = \mathbf{0}$ and $\mathcal{H}(\mathbf{0}) = \mathbf{0}$. Then, starting from $\mathbf{y}^{(0)}$ that is $\epsilon_{\mathbf{y}}$-distant from $\mathbf{y}^*$, using the step sizes $\frac{\gamma_l}{\|\mathbf{y}^{(0)}\|^2} \le \frac{2}{\beta}$, the iterates in (12) satisfy

$$\mathrm{dist}^2(\mathbf{y}^{(j)}, \mathbf{y}^*) \le \epsilon_{\mathbf{y}}^2 \left[\prod_{l=1}^{j}\left(1 - \frac{2\gamma_l}{\alpha\|\mathbf{y}^{(0)}\|^2}\right)\right]\|\mathbf{y}^*\|^2, \quad (36)$$

for $j \in [L]$, where $\alpha, \beta > 0$ are such that

$$\frac{4}{\alpha\beta} \le \left(\frac{\tilde{\mu}_{\mathcal{H}}}{\mu_{\mathcal{H}}}\right)^8 \left(\frac{h(\delta, \epsilon_{\mathbf{y}})}{c(\delta, \epsilon_{\mathbf{y}})}\right)^2. \quad (37)$$

Proof: See Section IX-D of the supplementary material. ■

This theorem unveils a number of important implications. Most notably, the concentration bound parameter δ is no longer the sole determinant of the recovery guarantee, as for the regime in (36) to be valid, several parameters must compositely satisfy the inequality $\delta_1 < 1$. Once this strict bound is violated, we no longer have a feasible α, β pair to guarantee the convergence in the encoded parameter space. This, in turn, requires

$$\delta < \left(\frac{\tilde{\mu}_H}{\mu_H}\right)^2 \frac{(1-\epsilon_{\boldsymbol{\rho}})(2-\epsilon_{\boldsymbol{\rho}})}{\sqrt{2}\omega(2+\epsilon_{\boldsymbol{\rho}})(2+\epsilon_{\mathbf{y}})}, \tag{38}$$

within our sufficient conditions of exact recovery. Furthermore, we can infer that $\tilde{\mu}_{\mathcal{H}}$, which is smaller than $\mu_{\mathcal{H}}$ by definition, should be away from 0 and $\epsilon_{\boldsymbol{\rho}}$ should be less than 1 as both of these constants affect the feasibility of the bound in (38).

B. Sketch of Proof for Theorem 1

Proof of the exact recovery guarantee in Theorem 1 depends on achieving an initial encoded image within a small neighborhood of the correct encoded unknown $\mathbf{y}^* \in \mathbb{Y}$. For our initialization scheme described in Section III and under the condition from (19), we have $\text{dist}^2(\boldsymbol{\rho}^{(0)}, \boldsymbol{\rho}^*) \leq \epsilon^2\|\boldsymbol{\rho}^*\|^2$ and

$$\text{dist}^2(\mathbf{y}^{(0)}, \mathbf{y}^*) \leq \epsilon_{\mathbf{y}}^2\|\mathbf{y}^*\|^2. \tag{39}$$

The inequality relation in (39) is derived in Section IX-C of the supplementary material. Our regularity condition states that for all $\mathbf{y} \in \mathbb{N}_{\epsilon_{\mathbf{y}}}(\mathbf{y}^*)$, $\mathcal{K}(\mathbf{y})$ satisfies the following inequality:

$$\text{Re}\left(\langle\nabla\mathcal{K}(\mathbf{y}), \mathbf{e}_{\mathbf{y}}\rangle\right) \geq \frac{1}{\alpha}\|\mathbf{e}_{\mathbf{y}}\|^2 + \frac{1}{\beta}\|\mathcal{K}(\mathbf{y})\|^2, \tag{40}$$

where $\mathbf{e}_{\mathbf{y}} = \mathbf{y} - \mathbf{y}^*$ and $\alpha, \beta > 0$. This ensures local strong convexity of $\mathcal{K}(\mathbf{y})$ within the $\epsilon_{\mathbf{y}}$ neighborhood of $\mathbf{y}^*$. Under (19) and (27), the regularity condition (40) is observed to be equivalent to

$$\frac{1}{\alpha\|\mathbf{y}^*\|^2} + \frac{1}{\beta}\mu_{\mathcal{H}}^8 c^2(\delta, \epsilon_{\mathbf{y}})\|\mathbf{y}^*\|^2 \leq h(\delta, \epsilon_{\mathbf{y}}). \tag{41}$$

Therefore, for (40) to be satisfied by $\mathcal{K}(\mathbf{y})$ for all $\mathbf{y} \in \mathbb{N}_{\epsilon_{\mathbf{y}}}(\mathbf{y}^*)$, the left hand side of (41) is required to be smaller than $h(\delta, \epsilon_{\mathbf{y}})$, which, in turn, leads to the condition in (37). Finally, by expanding $\|\mathbf{y}^{(l)} - \mathbf{y}^*\|$ using (12), and through (40) and the upper bound $\frac{2}{\beta}$ on the step sizes, we arrive at the result in (36).

C. Key Outcomes

a) Implications on the rate of convergence: By using fixed step sizes $\gamma \in \mathbb{R}^+$ for the L updates and by defining $\gamma' = \frac{\gamma}{\|\mathbf{y}^{(0)}\|^2} \leq \frac{2}{\beta}$, we observe from (36) that $\frac{2\gamma'}{\alpha}$ is a convergence rate related term where the convergence rate increases with an increase in its value. Furthermore, from Theorem 1, by using the upper bound $\frac{2}{\beta}$ on the step sizes, we can upper bound $\frac{2\gamma'}{\alpha}$ by $\frac{4}{\alpha\beta}$. Therefore, we can infer from (37) that $\frac{h^2(\delta)}{\mu_{\mathcal{H}}^8 c^2(\delta,\epsilon_{\mathbf{y}})}$ is essentially an upper bound on $\frac{2\gamma'}{\alpha}$. As long as $\tilde{\mu}_{\mathcal{H}}/\mu_{\mathcal{H}}$, $\epsilon_{\mathbf{y}}$, $\epsilon_{\boldsymbol{\rho}}$ and ω values are such that our modified upper bound on $\frac{2\gamma'}{\alpha}$ is larger than the one for the WF algorithm, our DL based approach will converge faster to the correct solution.

b) Conditions on the Lipschitz constants: From the definitions of $\epsilon_{\mathbf{y}}$ and $\epsilon_{\boldsymbol{\rho}}$, it is evident that with χ equal to $\tau \in \mathbb{R}^+$, upper bounding $\tau\mu_{\mathcal{H}}$ and $\tau\mu_{\mathcal{G}}\mu_{\mathcal{H}}^2\mu_{\mathcal{R}}(1+\epsilon)$ by $\xi_{\mathbf{y}} \in \mathbb{R}^+$ and $\xi_{\boldsymbol{\rho}} \in \mathbb{R}^+$, respectively, leads to the upper bound $\epsilon\xi_{\mathbf{y}}$ on $\epsilon_{\mathbf{y}}$ and $\epsilon\xi_{\boldsymbol{\rho}}$ on $\epsilon_{\boldsymbol{\rho}}$. It is shown in Section IX-E of the supplementary material that, $\tau\mu_{\mathcal{H}} \leq \xi_{\mathbf{y}} \leq 1$ and $\tau\mu_{\mathcal{G}}\mu_{\mathcal{H}}^2\mu_{\mathcal{R}}(1+\epsilon) \leq \xi_{\boldsymbol{\rho}} \leq 1$, if

$$\frac{(1-\tau\epsilon\mu_{\mathcal{H}})}{(1+\epsilon)} \leq \mu_{\mathcal{G}}\mu_{\mathcal{H}} \leq \min\left[2 - \frac{1}{\mu_{\mathcal{R}}}, \frac{\xi_{\boldsymbol{\rho}}}{\xi_{\mathbf{y}}}\right]\frac{1}{(1+\epsilon)}, \tag{42}$$

$$\mu_{\mathcal{H}} \leq \xi_{\mathbf{y}}/\tau, \tag{43}$$

$$\mu_{\mathcal{R}} \leq 1. \tag{44}$$

These bounds are sufficient for upper bounding $\epsilon_{\mathbf{y}}$ by $\epsilon\xi_{\mathbf{y}}$ and $\epsilon_{\boldsymbol{\rho}}$ by $\epsilon\xi_{\boldsymbol{\rho}}$. For a given τ and ω, if

$$\omega\left(\frac{\mu_{\mathcal{H}}}{\tilde{\mu}_{\mathcal{H}}}\right)^2 \frac{(2+\epsilon_{\boldsymbol{\rho}})(2+\epsilon_{\mathbf{y}})}{(1-\epsilon_{\boldsymbol{\rho}})(2-\epsilon_{\boldsymbol{\rho}})} \leq \frac{(2+\epsilon)}{\sqrt{(1-\epsilon)(2-\epsilon)}}, \tag{45}$$

then our exact recovery guarantee is valid over a larger range of δ compared to the WF algorithm.

c) Requirements on the $\mathbb{Y}$-domain: In the theorem statement, we assume that $\mathbb{Y}$ is an affine subset of $\mathbb{C}^{N_y}$. This assumption is made for mere convenience to deal the fact that the two eigenvectors of $\mathbf{E}_{\mathbf{y}} := \mathbf{y}\mathbf{y}^H - \mathbf{y}^*\mathbf{y}^{*H}$ are formed by normalized affine combinations of $\mathbf{y}$ and $\mathbf{y}^*$. This can be verified by following similar steps as the spectral analysis presented in Section IX-A of the supplementary material. For contractions in the parameter domain, the concentration property we imply via the $\mathcal{H}$-condition is required to hold over these eigenvectors, hence, we require that $\mathbb{Y}$ is an affine set, such that $\mathbf{u}_1 \in \mathbb{Y}$. Furthermore, this requirement can actually be relaxed to instead involve a *union of subspaces* model for $\mathbb{Y}$, since we merely need the union of pair-wise affine combinations of these elements $\mathbf{y}, \mathbf{y}^* \in \mathbb{Y}$.

This yields an interesting premise if the representations pursued for our image manifold are constrained to be *sparse* in the parameter space in $\mathbb{C}^{N_y}$. To this end, a $k-$sparsity constraint on representations results in the union of all 2 k-dimensional subspaces in $\mathbb{C}^{N_y}$ for $\mathbb{Y}$. Such a constraint however, must be enforced in the architecture via *projection* operators in the definition of the RNN-module. In our architecture and implementations, we do not provide any additional structure in $\mathbb{Y}$, and simply assume validity over all $\mathbb{C}^{N_y}$.

d) Spectral conditioning of $\mathcal{H}$: For convenience in presenting the theoretical results, we assume a global upper and lower Lipschitz property on $\mathcal{H}$ in (30). However, once an ϵ_y-neighborhood is guaranteed in the parameter space, it suffices that such a property is needed only locally over the neighborhood of a $\mathbf{y}^*$. To follow through with this relaxation, we need an additional spectral conditioning on $\mathcal{H}$, such that:

$$\tilde{\sigma}_{\mathcal{H}}\|\mathbf{y}\| \leq \|\mathcal{H}(\mathbf{y})\| \leq \sigma_{\mathcal{H}}\|\mathbf{y}\|, \tag{46}$$

for all $\mathbf{y} \in \mathbb{Y}$ where $\tilde{\sigma}_{\mathcal{H}}, \sigma_{\mathcal{H}} \in \mathbb{R}^+$. This is the basic premise of assuming that $\mathcal{H}$ is a *frame* over $\mathbb{Y}$. In this setting, the recovery guarantees promptly feature both ratio of μ_H and $\tilde{\mu}_H$, and the ratio of the frame coefficients, where the convergence bound becomes

$$\frac{4}{\alpha\beta} \leq \left(\frac{\tilde{\mu}_{\mathcal{H}}}{\mu_{\mathcal{H}}}\right)^4 \left(\frac{\tilde{\sigma}_{\mathcal{H}}}{\sigma_{\mathcal{H}}}\right)^4 \left(\frac{h(\delta, \epsilon_{\mathbf{y}})}{c(\delta, \epsilon_{\mathbf{y}})}\right)^2, \tag{47}$$

with the sufficient condition

$$\delta < \left(\frac{\tilde{\sigma}_{\mathcal{H}}\tilde{\mu}_{\mathcal{H}}}{\sigma_{\mathcal{H}}^2}\right)\frac{(1-\epsilon_{\boldsymbol{\rho}})(2-\epsilon_{\boldsymbol{\rho}})}{\sqrt{2}\omega(2+\epsilon_{\boldsymbol{\rho}})(2+\epsilon_{\mathbf{y}})}. \tag{48}$$

Most notably, with a linear $\mathcal{H}$, if (46) is satisfied over $\mathbb{C}^{N_y}$, all the ratios reduce to that of frame coefficients. This is highly relevant for the Gaussian linear encoder, which is the fundamental case that inspired our formulation under an arbitrary decoder. Namely, an over-determined Gaussian matrix satisfies the RIP over the whole domain in $\mathbb{C}^{N_y}$, with the RIP-constant $\delta_{\mathcal{H}} \in \mathbb{R}^+$ approaching 0 as M/N_y (i.e., the oversampling factor) grows, which increasingly well-conditions the problem, consistent with the statistical theory of phase retrieval.

VI. Training

A. Implementation of Lipschitz Constant Bounds

For our training set $\mathbb{D}$, let the intensity measurement vector and the associated ground truth image for the t^{th} sample, where $t \in [T]$, be denoted by $\mathbf{d}_t$ and $\boldsymbol{\rho}_t^*$, respectively. Training loss is computed as the average ℓ_2-norm difference between the estimated and the ground truth images. Moreover, since the image estimation $\boldsymbol{\rho}_t^{(l)}$, calculated as $\mathcal{H}(\mathbf{y}_t^{(l)})$ at the l^{th} RNN stage, is expected to get gradually closer to $\boldsymbol{\rho}_t^*$ as l increases, an additional term is typically added to the training loss function that sums the average ℓ_2-norm differences between $\boldsymbol{\rho}_t^{(l)}$ and $\boldsymbol{\rho}^*$. Our training loss $c_{tr}(\mathbb{U})$ is defined as

$$c_{tr}(\mathbb{U}) = \frac{1}{T}\sum_{t=1}^{T}\left[\|\hat{\boldsymbol{\rho}}_t - \boldsymbol{\rho}_t^*\|^2 + \sum_{l=1}^{L}\eta_l\|\mathcal{H}(\mathbf{y}_t^{(l-1)}) - \boldsymbol{\rho}_t^*\|^2\right] + c_0(\mathbb{U}). \tag{49}$$

$\eta_l \in \mathbb{R}^+$, where $l \in [L]$, is a constant, $\mathbb{U}$ denotes the set of parameters of the overall imaging network, and $c_0(\mathbb{U})$ is used to impose desirable properties on the trained networks. We set $c_0(\mathbb{U})$ as the sum of $c_i(\mathbb{U})$, where $i \in [4]$, and define $c_i(\mathbb{U})$ in the following discussion.

For imposing the property that $\mathcal{G}(\mathbf{0}) = \mathbf{0}$ and $\mathcal{H}(\mathbf{0}) = \mathbf{0}$, $c_1(\mathbb{U})$ can be set as $\eta_1(\|\mathcal{G}(\boldsymbol{\rho})|_{\boldsymbol{\rho}=\mathbf{0}}\|^2 + \|\mathcal{H}(\mathbf{y})|_{\mathbf{y}=\mathbf{0}}\|^2)$ where $\eta_1 \in \mathbb{R}^+$. In order to impose a specific Lipschitz constant value on the RNN, we define $c_2(\mathbb{U})$ as follows:

$$c_2(\mathbb{U}) = \eta_2\left(\max_{t_1,t_2\in[T]}\frac{\|\mathcal{R}(\mathbf{y}_{t_1}^{(0)}) - \mathcal{R}(\mathbf{y}_{t_2}^{(0)})\|}{\|\mathbf{y}_{t_1}^{(0)} - \mathbf{y}_{t_2}^{(0)}\|} - \mu_{\mathcal{R}}\right)^2, \tag{50}$$

where $\eta_2 \in \mathbb{R}^+$. The Lipschitz constants of $\mathcal{G}$ and $\mathcal{H}$ can be set to specific values using a similar approach as [49] by first setting $c_3(\mathbb{U})$ and $c_4(\mathbb{U})$ equal to $\eta_3\sum_{j=1}^{J}(\sigma(\mathbf{U}_j) - \mu_{\mathcal{G}}^j)^2$ and $\eta_4\sum_{k=1}^{K}(\sigma(\mathbf{W}_k) - \mu_{\mathcal{H}}^k)^2$, respectively, where $\eta_3, \eta_4 \in \mathbb{R}^+$, $\prod_{j=1}^{J}\mu_{\mathcal{G}}^j = \mu_{\mathcal{G}}$ and $\prod_{k=1}^{K}\mu_{\mathcal{H}}^k = \mu_{\mathcal{H}}$. $\sigma(.)$ and $\mathbf{U}_j$ are defined in Subsection III-B. $\mathbf{W}_k \in \mathbb{C}^{Q_k \times Q_{k-1}}$ is the weight matrix at the k^{th} layer of a similar $\mathcal{H}$ architecture as the one presented for $\mathcal{G}$ in Subsection III-B, where $k \in [K]$, $Q_0 = N_y$ and $Q_K = N$. While using the stochastic gradient descent to minimize $c_{tr}(\mathbf{W})$, in order to calculate the gradients of $c_3(\mathbb{U})$ and $c_4(\mathbb{U})$, we need to estimate the leading eigenvectors of the different weight matrices of $\mathcal{G}$ and $\mathcal{H}$, respectively. A power method is implemented in [49] where the leading eigenvectors estimated during one training update is reused as the initial vectors for the next update, for which the gradient of $c_{tr}(\mathbb{U})$ is calculated using a different mini-batch from the training set.

B. Computational Complexity

Computational complexity of our approach depends on the number of RNN stages L as well as the network architectures of $\mathcal{G}$ and $\mathcal{H}$. For linear activation functions for $\mathcal{G}$ and $\mathcal{H}$, forward propagations through these networks require $\sum_{j=1}^{J}P_jP_{j-1}$ and $\sum_{k=1}^{K}Q_kQ_{k-1}$ floating-point operations (FLOP), respectively. For ReLU activation functions and assuming that each comparison operation requires a single FLOP, an additional $\sum_{j=1}^{J-1}P_j + \sum_{k=1}^{K-1}Q_k + N_y + N$ FLOPs are carried out. The output of the $\mathcal{H}$ network is required to be calculated $L+1$ times. For the initial encoded image, we calculate the leading eigenvector of $\mathbf{Y}$, defined in (6), using the power method, and it incurs $O(N^3)$ computational cost. Calculating $\mathbf{F}\mathcal{H}(\mathbf{y}^{(l)})$ and then $\mathcal{F}(\mathcal{H}(\mathbf{y}^{(l)})\mathcal{H}(\mathbf{y}^{(l)})^H)$ requires $\mathcal{O}(MN) + \mathcal{O}(M)$ FLOPS in total. From $\mathcal{F}(\mathcal{H}(\mathbf{y}^{(l)})\mathcal{H}(\mathbf{y}^{(l)})^H)$, calculating $\frac{1}{M}\mathcal{F}^H(\mathbf{e})\mathcal{H}(\mathbf{y}^{(l)})$ takes another $\mathcal{O}(MN) + \mathcal{O}(M)$ operations. The error related term $\mathbf{e}$ is defined in Subsection III-A after (13). $\mathcal{H}(\mathbf{y}^{(l)})$ and its gradient $\nabla\mathcal{H}(\mathbf{y})|_{\mathbf{y}=\mathbf{y}^{(l)}}$ have updated values at each RNN stage, and the gradient is multiplied by an N length vector requiring an additional $\mathcal{O}(NN_y)$ FLOPS per iteration. With ReLU activation functions, $\mathcal{H}(\mathbf{y}^{(l)})$ calculation requires $\sum_{k=1}^{K}Q_kQ_{k-1} + \sum_{k=1}^{K-1}Q_k + N$ FLOPs. For calculating the gradient, the derivatives of the non-linear function require $N + \sum_{k=1}^{K-1}Q_k$ comparisons while the matrix multiplication part requires $\sum_{k=1}^{K}Q_kQ_{k-1} + \sum_{k=0}^{K-2}Q_kQ_{k+1}Q_{k+2}$ additional FLOPs. M is typically some constant multiple of N, where the constant is significantly smaller than N. If the value of Q_k, for $k \in [K-1]$, are in the order of N, then the computational complexity increases to $\mathcal{O}(N^3)$ per iteration. For this case, if the number of RNN stages L is significantly less than N, then the overall complexity remains $\mathcal{O}(N^3)$, similar to the generalized WF for interferometric inversion approach in [32]. On the other hand, for achieving an accuracy level of $\epsilon_{WF} \in \mathbb{R}^+$, the computational cost of the WF approach is $\mathcal{O}(N^2\log N\log(\frac{1}{\epsilon_{WF}}))$ [2].

VII. Numerical Simulations

In this section, we demonstrate the feasibility of our DL-based phaseless imaging approach through the training and subsequent performance evaluations on a number of real and simulated datasets, with measurement geometries of both experimental and practical interest. The main objectives of our numerical simulations are the following:

1) Demonstrating the reconstruction performance of our approach on both real and synthesized datasets, and comparing with the reconstruction results obtained using the WF algorithm [2], [8] and comparable DL-based state-of-the-art phaseless imaging methods, in order to highlight the

relative advantages of our approach over a range of image sets.

2) Numerically verifying the robustness of our approach under additive noise on the intensity measurements for relatively low $\frac{M}{N}$ values.

3) Numerically verifying a number of theoretical observations and insights presented in Section IV. These include showing the improved accuracy of the initial encoded image, resulting from the inclusion of $\mathcal{G}$, compared to the accuracy of the spectral estimation, observing the sample complexity improvement compared to the WF algorithm [2], [8] as well as other DL based approaches, and observing the necessity of having ample training set sizes for $\mathcal{H}$ to appropriately model various image classes of interest.

We adopt the normalized mean squared error (MSE) as the figure of merit throughout this section, and it is defined as $\text{MSE} = \frac{1}{T_s}\sum_{t=1}^{T_s}\|\hat{\boldsymbol{\rho}}_t - \boldsymbol{\rho}_t^*\|^2/\|\boldsymbol{\rho}_t^*\|^2$. T_s is the number of samples in the test set, $\mathbb{D}_{test}$, and $\hat{\boldsymbol{\rho}}_t$ and $\boldsymbol{\rho}_t^*$ denote the reconstructed and the corresponding ground truth images, respectively, for the t^{th} sample of $\mathbb{D}_{test}$.

A. Dataset Descriptions

In this subsection, we introduce three image sets, and the associated deterministic forward maps that results from three different data acquisition geometries.

1) MNIST Dataset: The first image set that we consider in this paper is MNIST, which is a publicly available dataset of handwritten digits. Each image has a dimension of 28×28 pixels, and depicts one of the 10 digits. We randomly select 10000 samples, with 1000 samples for each digit, as the training dataset, and another randomly selected 100 images, 10 for each digit, constitute the test set. For the forward mapping matrix, we use the one available with the publicly available dataset from [50] for the 40×40 pixels imaging scenario. This dataset considers a multiple scattering transmission environment with phaseless measurements, and the forward map is recovered using the prVAMP based double phase retrieval approach. Since our images have a lower pixel count, we consider the first 784 columns of this matrix to form our forward map $\mathbf{F}$, and discard the phases of the $\mathbf{F}\boldsymbol{\rho}_t^*$ values to form the phaseless measurements for the images in the MNIST dataset. The number of rows of $\mathbf{F}$, which is the number of total measurements M, is varied for experimentation purposes, and for each case, we consider the first M rows of $\mathbf{F}$.

2) Simulated Synthetic Aperture Dataset: The second dataset is selected with the goal of showing a scenario where our approach is applicable in a practical setting with a deterministic forward map. We apply our method for synthetic aperture imaging [51] from simulated measurement under Born approximation. Each scene being imaged has a dimension of $500\text{ m} \times 500\text{ m}$ and is reconstructed as a 14×14 pixels image. There is a single square object located at a random location within the area being imaged, and the background varies from scene to scene. The number of samples in the training and test sets are 9950 and 50, respectively. We consider a mono-static data-collection strategy, with the transmitter-receiver trajectory along a circular path at 7 km height and at a radius of 10 km. Total number of measurements is set equal to the number of unknowns, i.e., $M = 196$, and additive Gaussian noise of zero mean and different variances is assumed to be present in the measured intensity values of the received signal. A schematic diagram of the associated data collection geometry is shown in Fig. 3.

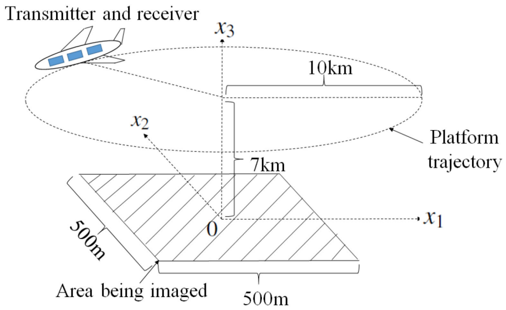


Fig. 3. Data collection geometry for the synthetic aperture imaging.

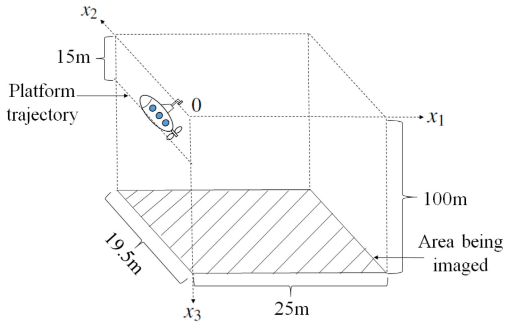


Fig. 4. Data collection geometry for PCSWAT dataset for SAS imaging.

3) PCSWAT Generated SAS Dataset: For the third dataset, we consider a PCSWAT 10 software generated simulated dataset for synthetic aperture sonar (SAS) imaging of underwater scenes. PCSWAT is a tool-set developed for simulating high-fidelity SAS data [52]. It offers a selection of realistic targets and underwater surface types, and allows the incorporation of varying sound-velocity profiles, marine life property, wind speed etc. For the samples in our training and test sets, we consider that the background medium is composed of sandy gravel, and there is sparse marine life present in the medium. Each scene contains a single hemi-spherically end-capped cylinder of varying length and fixed radius located at a random location on the scene along a random orientation. Each area being imaged has a dimension of 19 m ×25 m and it is reconstructed as a 22×31 pixels image. The number of samples in the training and test set are 800 and 5, respectively. We consider a 2D environment, and the vehicle and the water depth are set to 15 m and 100 m, respectively. The center frequency and the bandwidth of the transducers mounted on the moving vehicle are set equal to 120 kHz and 30 kHz, respectively. The data collection geometry for the SAS operation simulated via PCSWAT is shown in Fig. 4.

B. DL Architectures and Reconstruction Results

The quality of the reconstructed images is heavily dependent on the $\mathcal{G}$ and $\mathcal{H}$ network architectures. For evaluating our numerical results for the MNIST dataset, we consider the following network model: the number of RNN stages is set to 10; for $\mathcal{G}$, we use a 5 layer CNN model with $leaky\_relu(.)$ activation functions from the tensorflow library, and output dimensions $24 \times 24 \times 4$, $20 \times 20 \times 16$, $16 \times 16 \times 16$, $12 \times 12 \times 4$ and $8 \times 8 \times 1$ for the 5 consecutive layers; for $\mathcal{H}$, we have used a 5 layer ANN architecture with $relu(.)$ activation functions, and output vector lengths of 64, 64, 64, 100 and 784, respectively. Additionally, since the maximum value of each MNIST image is 255, we consider this as prior information while applying our DL-based approach as well as the other methods that we evaluate for comparison. This is performed by first normalizing the set of intensity vectors by 255^2, and then adding the term $\frac{1}{T}\sum_{t=1}^{T}(\max_{n\in[N]} \hat{\boldsymbol{\rho}}_t(n) - 1)^2$, where $\hat{\boldsymbol{\rho}}_t(n)$ denotes the n^{th} element of $\hat{\boldsymbol{\rho}}_t$, to the training loss functions of the end-to-end DL-based methods. On the other hand, for the iterative methods including the WF algorithm, we normalize the updated image estimation at each iteration so that the maximum pixel value equals to 1.

The network models implemented for the synthetic aperture and the PCSWAT generated SAS datasets are kept similar as the ones used for MNIST. For synthetic aperture imaging, the number of filters in $\mathcal{G}$ network layers are 8, 12, 12, 8 and 1, respectively, while the output vector lengths of the 5 consecutive layers of $\mathcal{H}$ are 81, 85, 90, 100 and 196, respectively. For the SAS dataset, the number of filters used in the $\mathcal{G}$ network layers are the same, except, in this case, the encoded image dimension N_y is set to 64. The output vector lengths for the 5 consecutive $\mathcal{H}$ layers are 64, 81, 100, 150 and 200, respectively. We used ADAM optimizer for training, with a learning rate equal to 10^{-5}, and batch sizes equal to 100, 50 and 5 for the MNIST, synthetic aperture and the PCSWAT datasets, respectively.

For the remainder of this section, we refer to our approach by DL-WF. We note that while comparing with other DL-based state-of-the-art phaseless imaging approaches, we exclude comparisons with the generative prior based methods [9]–[12], as they require us to separately train a GAN using a large amount of images from comparable image classes. One of the motivations of our approach is to avoid the cumbersome GAN network training, and instead adopt an end-to-end training strategy that uses sample images and the corresponding intensity measurements. Additionally, although our theoretical results do not guarantee performance improvement over the generative prior based methods, there are several advantages of our exact recovery guarantee compared to the theoretical results derived in [9], [11] as summarized in Section IV. In this section, we instead include reconstruction results from two state-of-the-art DL-based approaches with comparable training complexities, namely, UPR [29] and prDeep [30]. UPR [29] uses an end-to-end training scheme, with similar training dataset requirement as DL-WF. On the other hand, prDeep is a regularization by denoising [53] type approach for phaseless imaging, and it implements a DnCNN [54] for denoising. We have used a 17 layer DnCNN network with a similar architecture as the one presented in [54], where the number of channels at each intermediate layer is 64. Instead of patch extraction, due to the relatively small image dimensions under consideration here, we apply the entire image as input to the denoising network. For additional implementation details for UPR and prDeep, we followed the various hyper-parameter values suggested in [29] and [30], respectively.

Example reconstructed images using our DL-based method along with the reconstructed images using the WF algorithm, prDeep and UPR are shown in Figs. 5, 6 and 7 for the MNIST, synthetic aperture and PCSWAT datasets, respectively. The number of training samples used for these three datasets are 10000, 9950 and 800, respectively. For all three cases, we observe that our DL-WF approach yields significant improvement in the estimated image accuracies compared to the WF algorithm, and the DL-based UPR and prDeep methods.

Despite this improvement, the reconstructed images produced by our approach still retain visual artifacts. There are two key contributors to this end. Firstly, Section V discusses exact recovery with linear convergence for elements representable in the range of $\mathcal{H}$. For the particular training dataset and the optimization algorithm used for training, whether we can estimate an $\mathcal{H}$ with the properties specified in the theory of exact recovery is an additional aspect that contributes to empirically observing such guarantees. Secondly, even under the validity of these assumptions, observing exact recovery still potentially requires many iterations of gradient updates in the lower dimensional encoded space given the ill-posed problem settings under consideration. The architecture is however limited by the number of layers in the RNN unit, hence convergence to the true solution is not necessarily observed. Accordingly in Fig. 12(c), we demonstrate the expected decaying trend in average reconstruction error as the number of RNN stages are increased. Furthermore, despite these limitations, the improvements in the reconstruction quality that our approach offers over the state-of-the-art phaseless imaging methods is still quite significant.

C. Effect of Sample Complexity

In order to observe the effect of sample complexity on our approach, MSE values for the MNIST test dataset are plotted versus the $\frac{M}{N}$ ratios in Fig. 8. It is observed that, for each of the $\frac{M}{N}$ ratios, our DL-based approach performs better compared to the WF algorithm, prDeep and UPR. Additionally, as expected intuitively, we observe reduced MSE values as M is increased for a fixed image dimension. In Fig. 9(a), 9(b) and 9(c), we consider the case, where the number of measurements M is fixed while the encoded image dimension N_y is varied, and we plot the MSE values versus $\sqrt{N_y}$ for the MNIST, synthetic aperture and the PCSWAT datasets, respectively. For all three cases, we observe reduced MSE values with increasing N_y. Our observation from Fig. 9 implies that the reconstruction in the encoded image space $\mathbb{Y}$, reveals a latent dimension of the images smaller than the number of unknowns. This indicates that, compared to the WF algorithm and the state-of-the-art DL-based methods, our approach has lower sample complexity requirement since

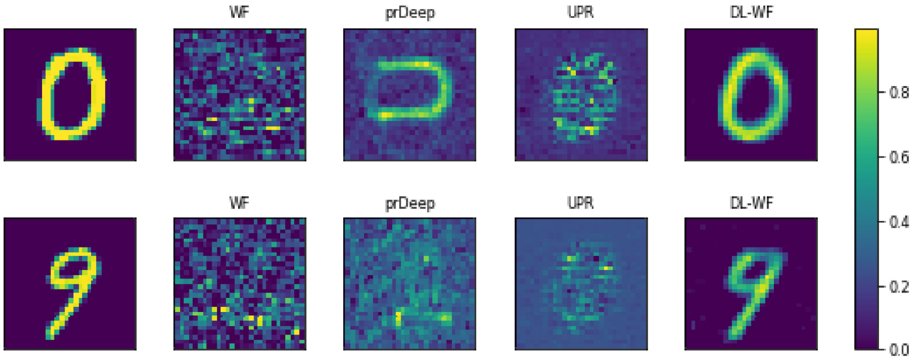


Fig. 5. First column includes the original unknown images of dimension 28 × 28 pixels. For $M = 0.5\ N$ and 10000 training samples, the reconstructed images using the WF algorithm [2] with 5000 iterations are shown in the second column. Corresponding estimated images using the prDeep [30] and the UPR [29] approaches are included in the third and fourth columns, respectively. The last column shows the estimated images using our method with 10 RNN stages.

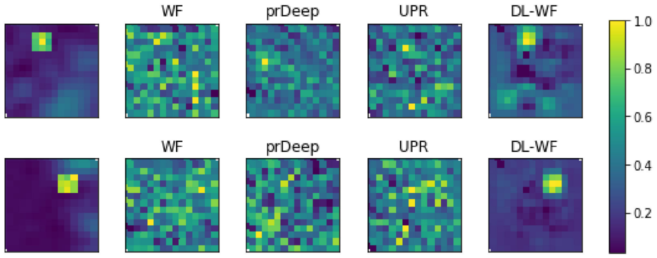


Fig. 6. Image reconstruction results for the simulated synthetic aperture dataset with 14 × 14 pixel images and SNR = 10dB. For $M = N$ and 9950 training samples, the five columns show the original images, and the reconstructed images using the WF algorithm [2] with 5000 iterations, prDeep approach [30], UPR approach [29] and our DL-WF approach with 10 RNN stages, respectively.

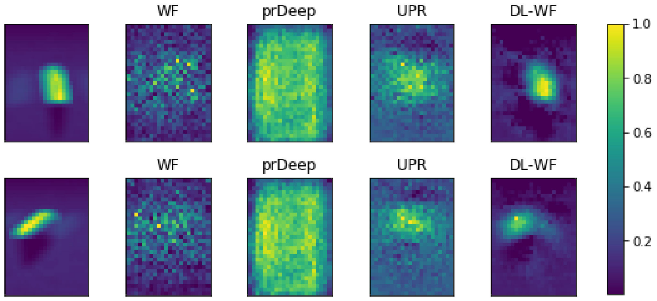


Fig. 7. Image reconstruction results for the PCSWAT dataset with 22 × 31 pixel underwater scenes and SAS measurements. For $M = 930$ and 800 training samples, the five columns show the original images, and the estimated images using the WF algorithm [2] with 5000 iterations, prDeep approach [30], UPR approach [29] and our DL-based approach with 10 RNN stages, respectively.

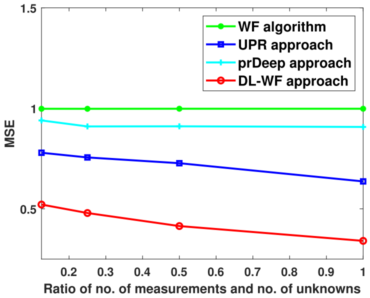


Fig. 8. MSE versus $\frac{M}{N}$ for the samples in the MNIST test set.

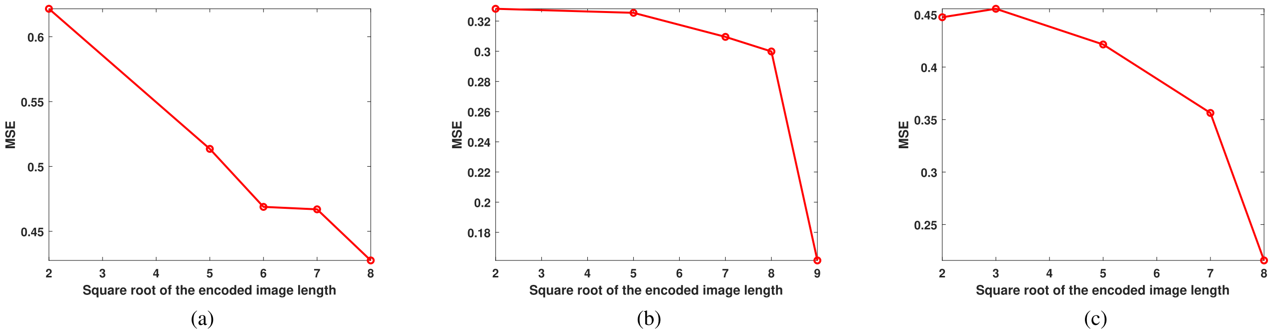
it searches over a reduced space for the unknown image, as supported by the corresponding MSE values in Fig. 8.

D. Effect of the Number of Training Samples

Another important criteria is the necessity of having an adequate number of training samples for effective image reconstruction at the decoding network output. In Fig. 10, we plot the MSE values for the test set versus the number of training set sizes for $M = 1.36\ N$ for the PCSWAT dataset. We consider two cases with the same $\mathcal{G}$ and $\mathcal{H}$ network architectures, and the number of RNN stages equal to 5 and 10. As expected, we observe for both cases that an increasing training set size helps $\mathcal{H}$ to capture the underlying image prior more effectively as long as the $\mathcal{H}$ architecture has sufficient capacity. Additionally, it helps the encoder to learn a better encoded image space while simultaneously attaining improved RNN parameter values, which translates into lower MSE values at similar stages of the training process. As for the two curves corresponding to the different numbers of RNN layers, we observe that as long as the overall imaging network is sufficiently trained, which is represented by the last points on both curves, increasing the number of RNN layers helps improve the reconstruction accuracies.

E. Accuracy of the Initial Value

In order to observe the effect of $\mathcal{G}$ on the initialization accuracy and to indirectly verify the observation from (39), we consider three separate mean initialization error related terms for the samples in the test sets, namely, d_1, d_2 and d_3. We define d_1, d_2 and d_3 as $d_1 = \frac{1}{T_s}\sum_{t=1}^{T_s}\|\boldsymbol{\rho}_t^{(0)} - \boldsymbol{\rho}_t^*\|^2/\|\boldsymbol{\rho}_t^*\|^2$, $d_2 = \frac{1}{T_s}\sum_{t=1}^{T_s}\|\mathcal{H}(\mathbf{y}_t^{(0)}) - \boldsymbol{\rho}_t^*\|^2/\|\boldsymbol{\rho}_t^*\|^2$ and $d_3 = \frac{1}{T_s}\sum_{t=1}^{T_s}\|\mathbf{y}_t^{(0)} - \mathbf{y}_t^{(L)}\|^2/\|\mathbf{y}_t^{(L)}\|^2$. A more accurate calculation of the initialization error for the encoded image space, d_3, requires $\mathbf{y}_t^{(L)}$ to be replaced by $\mathbf{y}_t^*$. When the $\mathcal{G}$ and $\mathcal{H}$ network architectures, and the numbers of training samples are set as described in Subsection VII-B, and the number of RNN stages is set to 10, we observe that for the three datasets, the three initialization error related terms have the following values: for MNIST with $M = 0.5\ N$, $d_1 = 209.344$, $d_2 = 0.999989$ and $d_3 = 0.000145729$; for the synthetic aperture dataset with $M = N$, $d_1 = 2.02657$,



(a) (b) (c)

Fig. 9. MSE versus $\sqrt{N_y}$ for fixed M/N ratio for the samples in the test set (a) for the MNIST dataset with $M = 0.5\ N$, (b) for the synthetic aperture dataset with $M = N$, and (c) for the PCSWAT dataset with $M = 1.36\ N$, and the number of RNN stages $L = 5$.

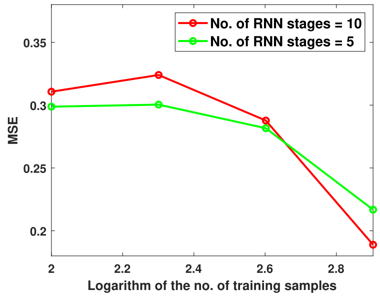


Fig. 10. MSE versus the training set sizes for the PCSWAT dataset with $M = 1.36\ N$.

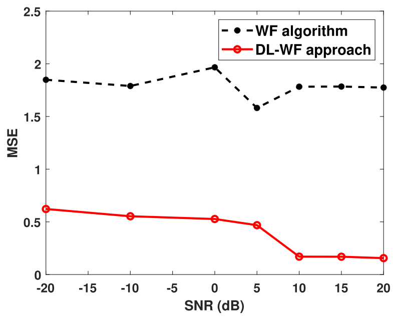


Fig. 11. MSE versus SNR (dB) for $M = N$ for the samples in the synthetic aperture test dataset.

$d_2 = 0.300721$ and $d_3 = 0.00103663$; and for the PCSWAT dataset with $M = 1.36\ N$, $d_1 = 1.49407$, $d_2 = 0.525142$ and $d_3 = 0.00129634$. In all three cases, we observe that the trained networks produce significantly reduced initialization errors for the encoded image space compared to the ones for the original image space.

F. Effect of SNR of the Intensity Measurements

The effect of varying SNR values, resulting from the different levels of noise detected at the receiving sensors along with the intensity values of the reflected signals, on the accuracies of the reconstructed images for the synthetic aperture dataset is shown in Fig. 11. We compare these values to the corresponding image reconstruction accuracies for the WF algorithm. With increasing SNR, we observe some reduction in the MSE values, calculated after a fixed number of training updates. For each case, our DL method is observed to significantly outperform the WF algorithm.

G. Effect of $\mathcal{G}$ and $\mathcal{H}$ Architectures and No. Of RNN Layers

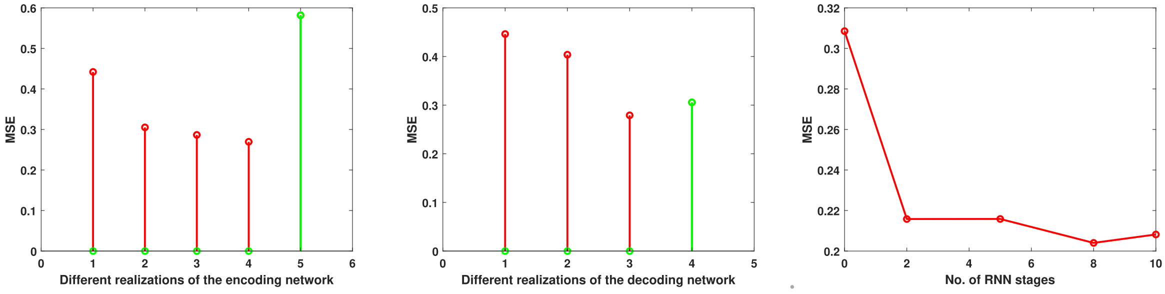
In this subsection, we show the effects of the encoding and decoding network architectures, and the number of RNN stages on the performance of our approach. We consider the PCSWAT dataset for this purpose, and while evaluating the effect of each of these criteria, for example the $\mathcal{G}$ network architecture, we keep the remaining elements, i.e. the $\mathcal{H}$ network architecture and the number of RNN layers, unchanged. In Fig. 12(a) and 12(b), each point along the x-axis, corresponds to one realization of the $\mathcal{G}$ and $\mathcal{H}$ networks, respectively. The number of parameters for these different architectures increase from left to right, and the linear architectures for each case are indicated by the last points, where the corresponding linear networks have the maximum number of trainable parameters. These figures provide the empirical observation that, the MSE value reduces with $\mathcal{G}$ and $\mathcal{H}$ network architectures with increasing number of parameters, and a linear encoder is more detrimental than a linear decoder. Finally, Fig. 12(c) verifies that with an increasing number of RNN layers, we can improve our reconstruction quality.

H. Comparison With DL Methods for Fourier Measurements

In this subsection, we compared our approach to prDeep and UPR, under Fourier measurement models. We use the images from the PCSWAT dataset along with two cases of the Fourier forward map, where the number of measurements equal to 1.5 times and 2 times the number of unknowns. For our DL-based approach, we adopt a 5 layer RNN, with $\mathcal{G}$ and $\mathcal{H}$ network architectures as presented in Subsection VII-B for the PCSWAT dataset. Example reconstructed images resulting from the three approaches are shown in Fig. 13. For both values of $\frac{M}{N}$, we observe that our approach outperforms the state-of-the-art DL-based methods under Fourier measurements.

I. Comparison of Computation Time

Computation times during the inversion phases for our approach and the UPR method are dominated by the time required



(a) (b) (c)

Fig. 12. MSE for the PCSWAT dataset using different (a) $\mathcal{G}$ and (b) $\mathcal{H}$ network architectures and 5 RNN layers. The last points on the x-axis correspond to the linear encoding and decoding network architectures in (a) and (b), respectively. (c) shows the MSE values obtained by using different numbers of RNN stages for fixed $\mathcal{G}$ and $\mathcal{H}$ network architectures.

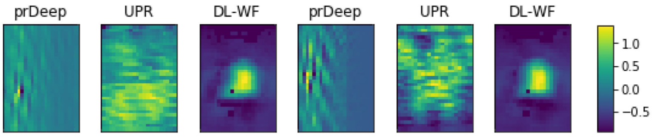


Fig. 13. Reconstructed images for the samples in the PCSWAT test imageset from the corresponding Fourier measurements using different DL-based methods. For the ground truth image in the top left corner of Fig. 7, the first three figures show the reconstructed images using prDeep, UPR and DL-WF, respectively, for $M = 1.5\ N$; next three figures show the corresponding images for $M = 2\ N$.

to compute the spectral initialization output. On the other hand, for the prDeep approach, significant computational time is involved both for training the denoising network and the inversion phase. Over all three datasets, we observe that the 20-layer RNN network for the UPR approach has the lowest computational time, followed by our DL-WF approach, while the WF algorithm and the prDeep approach require the highest computational time during the inversion phase. As an example, for the PCSWAT dataset, the average computational times required in the inversion phases of the WF algorithm, prDeep, UPR and DL-WF with $L = 5$ are 3.8815, 9.8472, 0.0147, and 0.0148 minutes, respectively. For the same dataset, the training time for the UPR and the DL-WF approaches is approximately 2-3 days, while the training time for the denoising network of the prDeep approach is approximately a few hours.

VIII. Conclusion

In this paper, we introduced a DL-based phaseless imaging approach that incorporates an RNN with DL-based encoding-decoding stages, and determined sufficient conditions for exact recovery guarantee. Our theoretical results show that, depending on the Lipschitz constants of the encoding and the decoding networks, it is possible to achieve improved convergence rate as compared to the WF algorithm [2]. Additionally, the valid range of forward maps for which the exact recovery guarantee holds is less restrictive than those sufficient conditions introduced in earlier works [8], [9], [11]. Desired spectral property of the decoder for the feasibility of our recovery guarantee reveals that our theoretical results are consistent with the observations for the case with linear Gaussian generative priors and forward maps with i.i.d. Gaussian distributed elements. Our numerical simulations show the advantages of our approach, under low sample complexity regimes and deterministic forward maps, over the WF algorithm as well as the existing DL-based methods. In future work, we will consider extensions to take into account partially known forward maps which relates to a multiple scattering within extended objects scenario in practical remote sensing applications. Additionally, we will pursue improvements to our approach with deep equilibrium architectures [55] to facilitate more iterations on the lower dimensional encoded space towards higher accuracy in reconstructions.

References

[1] M. Soltanolkotabi, “Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization,” *IEEE Trans. Inf. Theory*, vol. 65, no. 4, pp. 2374–2400, Apr. 2019.

[2] E. J. Candes, X. Li, and M. Soltanolkotabi, “Phase retrieval via wirtinger flow: Theory and algorithms,” *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1985–2007, Apr. 2015.

[3] K. Jaganathan, Y. Eldar, and B. Hassibi, “Phase retrieval with masks using convex optimization,” in *Proc. IEEE Int. Symp. Inf. Theory*, 2015, pp. 1655–1659.

[4] T. T. Cai, X. Li, and Z. Ma, “Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow,” *Ann. Statist.*, vol. 44, no. 5, pp. 2221–2251, 2016.

[5] F. Wu and P. Rebeschini, “Hadamard wirtinger flow for sparse phase retrieval,” in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 982–990.

[6] Y. Chen and E. Candes, “Solving random quadratic systems of equations is nearly as easy as solving linear systems,” *Adv. Neural Inf. Process. Syst.*, vol. 28, pp. 739–747, 2015.

[7] Z. Yuan, H. Wang, and Q. Wang, “Phase retrieval via sparse wirtinger flow,” *J. Comput. Appl. Math.*, vol. 355, pp. 162–173, 2019.

[8] B. Yonel and B. Yazici, “A deterministic theory for exact non-convex phase retrieval,” *IEEE Trans. Signal Process.*, vol. 68, pp. 4612–4626, 2020.

[9] P. Hand, O. Leong, and V. Voroninski, “Phase retrieval under a generative prior,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 9136–9146.

[10] F. Shamshad and A. Ahmed, “Compressed sensing-based robust phase retrieval via deep generative priors,” *IEEE Sens. J.*, vol. 21, no. 2, pp. 2286–2298, 2021.

[11] P. Hand, O. Leong, and V. Voroninski, “Compressive phase retrieval: Optimal sample complexity with deep generative priors,” 2020, *arXiv:2008.10579*.

[12] F. Shamshad and A. Ahmed, “Compressed sensing-based robust phase retrieval via deep generative priors,” *IEEE Sensors J.*, vol. 21, no. 2, pp. 2286–2298, Jan. 2021.

[13] G. Jagatap and C. Hegde, “Algorithmic guarantees for inverse imaging with untrained network priors,” *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 14832–14842, 2019.

[14] H. Zhang, Y. Liang, and Y. Chi, “A nonconvex approach for phase retrieval: Reshaped wirtinger flow and incremental algorithms,” *J. Mach. Learn. Res.*, vol. 18, no. 141, pp. 1–35, 2017. [Online]. Available: http://jmlr.org/papers/v18/16-572.html

[15] E. J. Candes, T. Strohmer, and V. Voroninski, “Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming,” *Commun. Pure Appl. Math.*, vol. 66, no. 8, pp. 1241–1274, Aug. 2013.

[16] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski, “Phase retrieval via matrix completion,” *SIAM Rev.*, vol. 57, no. 2, pp. 225–251, 2015.

[17] H. Zhang, Y. Chi, and Y. Liang, “Median-truncated nonconvex approach for phase retrieval with outliers,” *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 7287–7310, Nov. 2018.

[18] Y. M. Lu and G. Li, “Phase transitions of spectral initialization for high-dimensional non-convex estimation,” *Inf. Inference: A J. IMA*, vol. 9, no. 3, pp. 507–541, Sep. 2020.

[19] W. Luo, W. Alghamdi, and Y. M. Lu, “Optimal spectral initialization for signal recovery with applications to phase retrieval,” *IEEE Trans. Signal Process.*, vol. 67, no. 9, pp. 2347–2356, May 2019.

[20] B. Yonel and B. Yazici, “A spectral estimation framework for phase retrieval via bregman divergence minimization,” *SIAM J. Imag. Sci.*, vol. 15, no. 2, pp. 491–520, 2022.

[21] R. Ghods, A. Lan, T. Goldstein, and C. Studer, “Linear spectral estimators and an application to phase retrieval,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1734–1743.

[22] G. Wang, G. B. Giannakis, and Y. C. Eldar, “Solving systems of random quadratic equations via truncated amplitude flow,” *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 773–794, Feb. 2018.

[23] D. Gilton, G. Ongie, and R. Willett, “Neumann networks for linear inverse problems in imaging,” *IEEE Trans. Comput. Imag.*, vol. 6, pp. 328–343, 2020.

[24] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Deep image prior,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9446–9454.

[25] R. Heckel and P. Hand, “Deep decoder: Concise image representations from untrained non-convolutional networks,” in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–14.

[26] V. Monga, Y. Li, and Y. C. Eldar, “Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing,” *IEEE Signal Process. Mag.*, vol. 38, no. 2, pp. 18–44, Mar. 2021.

[27] R. Hyder, Z. Cai, and M. S. Asif, “Solving phase retrieval with a learned reference,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 425–441.

[28] F. Zhang, X. Liu, C. Guo, S. Lin, J. Jiang, and X. Ji, “Physics-based iterative projection complex neural network for phase retrieval in lensless microscopy imaging,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10523–10531.

[29] N. Naimipour, S. Khobahi, and M. Soltanalian, “UPR: A model-driven architecture for deep phase retrieval,” in *Proc. IEEE 54th Asilomar Conf. Signals, Systems, Comput.*, 2020, pp. 205–209.

[30] C. Metzler, P. Schniter, A. Veeraraghavan, and R. Baraniuk, “prDeep: Robust phase retrieval with a flexible deep network,” in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, vol. 80, pp. 3501–3510.

[31] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 1–16.

[32] B. Yonel and B. Yazici, “A generalization of wirtinger flow for exact interferometric inversion,” *SIAM J. Imag. Sci.*, vol. 12, no. 4, pp. 2119–2164, 2019.

[33] G. S. Sammelmann, “Personal computer shallow water acoustic tool-set (PC SWAT) 7.0: Low frequency propagation and scattering,” Naval Surface Warfare Center, Panama City, Florida, Tech. Rep. CSS/TR-02/10, 2002.

[34] G. Wang, G. Giannakis, Y. Saad, and J. Chen, “Solving most systems of random quadratic equations,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1867–1877.

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

[36] P. Netrapalli, P. Jain, and S. Sanghavi, “Phase retrieval using alternating minimization,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2796–2804.

[37] D. R. Luke, “Relaxed averaged alternating reflections for diffraction imaging,” *Inverse Problems*, vol. 21, no. 1, 2004, Art. no. 37.

[38] A. H. Barnett, C. L. Epstein, L. F. Greengard, and J. F. Magland, “Geometry of the phase retrieval problem,” *Inverse Problems*, vol. 36, no. 9, 2020, Art. no. 094003.

[39] S. Bahmani and J. Romberg, “Phase retrieval meets statistical learning theory: A flexible convex relaxation,” in *Proc. Artif. Intell. Statist.*, 2017, pp. 252–260.

[40] T. Goldstein and C. Studer, “PhaseMax: Convex phase retrieval via basis pursuit,” *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2675–2689, Apr. 2018.

[41] P. Hand and V. Voroninski, “An elementary proof of convex phase retrieval in the natural parameter space via the linear program PhaseMax,” *Commun. Math. Sci.*, vol. 16, no. 7, pp. 2047–2051, 2018.

[42] E. Mason, B. Yonel, and B. Yazici, “Deep learning for radar,” in *Proc. IEEE Radar Conf.*, 2017, pp. 1703–1708.

[43] B. Yonel, E. Mason, and B. Yazıcı, “Deep learning for passive synthetic aperture radar,” *IEEE J. Sel. Top. Signal Process.*, vol. 12, no. 1, pp. 90–103, Feb. 2018.

[44] B. Yonel, E. Mason, and B. Yazici, “Deep learning for waveform estimation and imaging in passive radar,” *IET Radar, Sonar Navigation*, vol. 13, no. 6, pp. 915–926, 2019.

[45] S. Kazemi, B. Yonel, and B. Yazici, “Deep learning for direct automatic target recognition from SAR data,” in *Proc. IEEE Radar Conf.*, 2019, pp. 1–6.

[46] S. Kazemi and B. Yazici, “Deep learning for joint image reconstruction and segmentation for SAR,” in *Proc. IEEE Int. Radar Conf.*, 2020, pp. 839–843.

[47] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, “Spectral normalization for generative adversarial networks,” in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–26.

[48] K. Jaganathan, S. Oymak, and B. Hassibi, “Sparse phase retrieval: Uniqueness guarantees and recovery algorithms,” *IEEE Trans. Signal Process.*, vol. 65, no. 9, pp. 2402–2410, May 2017.

[49] Y. Yoshida and T. Miyato, “Spectral norm regularization for improving the generalizability of deep learning,” 2017, *arXiv:1705.10941*.

[50] C. A. Metzler, M. K. Sharma, S. Nagesh, R. G. Baraniuk, O. Cossairt, and A. Veeraraghavan, “Coherent inverse scattering via transmission matrices: Efficient phase retrieval algorithms and a public dataset,” in *Proc. IEEE Int. Conf. Comput. Photogr.*, 2017, pp. 1–16.

[51] J. K. Jao, “Theory of synthetic aperture radar imaging of a moving target,” *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 9, pp. 1984–1992, Sep. 2001.

[52] D. Marx, M. Nelson, E. Chang, W. Gillespie, A. Putney, and K. Warman, “An introduction to synthetic aperture sonar,” in *Proc. 10th IEEE Workshop Stat. Signal Array Process.*, 2000, pp. 717–721.

[53] Y. Romano, M. Elad, and P. Milanfar, “The little engine that could: Regularization by denoising (RED),” *SIAM J. Imag. Sci.*, vol. 10, no. 4, pp. 1804–1844, 2017.

[54] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, “Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising,” *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[55] D. Gilton, G. Ongie, and R. Willett, “Deep equilibrium architectures for inverse problems in imaging,” *IEEE Trans. Comput. Imag.*, vol. 7, pp. 1123–1133, Oct. 2021.

Samia Kazemi received the B.Sc. degree in electrical and electronics engineering from the Bangladesh University of Engineering and Technology, Dhaka, Bangladesh, in 2012, and the M.Sc. degree in electrical engineering in 2016 from Rensselaer Polytechnic Institute, Troy, NY, USA, where she is currently working toward the Ph.D. degree. Her research interests include deep learning, optimization methods for inverse problems in imaging, pattern recognition, and information theory.

Bariscan Yonel (Member, IEEE) received the B.Sc. degree in electrical engineering from Koc University, Istanbul, Turkey, in June 2015, and the Ph.D. degree in electrical engineering in December 2020 from Rensselaer Polytechnic Institute, Troy, NY, USA, where he is currently a Postdoctoral Research Associate with the Computational Imaging Group. His work focuses on theoretical guarantees and practical limitations for solving quadratic equations in high dimensional inference and wave-based imaging problems, using low rank matrix recovery theory and computationally efficient non-convex algorithms. His research interests include applications and performance analysis of machine learning, compressed sensing and optimization methods for inverse problems in imaging, and signal processing.

Birsen Yazici (Fellow Member, IEEE) received the B.S. degree in electrical engineering and mathematics from Bogazici University, Istanbul Turkey, in 1988 and the M.S. and Ph.D. degrees in mathematics and electrical engineering from Purdue University, West Lafayette, IN, USA, in 1990 and 1994, respectively. From September 1994 to 2000, she was a Research Engineer with General Electric Company Global Research Center, Schenectady, NY, USA. During her tenure in industry, she worked on radar, transportation, industrial and medical imaging systems. From 2001 to June 2003, she was an Assistant Professor with Drexel University, Philadelphia, PA, USA, Electrical and Computer Engineering Department. In 2003, she joined Rensselaer Polytechnic Institute where she is currently a Full Professor with the Department of Electrical, Computer and Systems Engineering and with the Department of Biomedical Engineering. Her research interests include statistical signal processing, inverse problems in imaging, image reconstruction, biomedical optics, radar and X-ray imaging. She was an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING from 2008 to 2012, IEEE TRANSACTIONS ON GEOSCIENCES AND REMOTE SENSING from 2014 to 2018, for *SIAM Journal on Imaging Science* from 2010 to 2014, and for *IEEE Transactions on Computational Imaging* from 2017 to 2020. She is currently a Distinguished Lecturer of the IEEE Aerospace and Electronics Systems Society. She was the recipient of the Rensselaer Polytechnic Institute 2007 and 2013 School of Engineering Research Excellence Awards. She holds 11 U.S. patents.