IEEE Open Journal of Signal Processing

Received 10 October 2023; revised 21 February 2024; accepted 27 February 2024. Date of publication 8 March 2024; date of current version 23 April 2024. The review of this article was arranged by Associate Editor Esteban Vera.

Digital Object Identifier 10.1109/OJSP.2024.3375276

Short Paper

PtychoDV: Vision Transformer-Based Deep Unrolling Network for Ptychographic Image Reconstruction

WEIJIE GAN [1] (Student Member, IEEE), QIUCHEN ZHAI [2] (Graduate Student Member, IEEE), MICHAEL T. MCCANN [3] (Member, IEEE), CRISTINA GARCIA CARDONA [3], ULUGBEK S. KAMILOV [1,4] (Senior Member, IEEE), AND BRENDT WOHLBERG [3] (Fellow, IEEE)

[1]Department of Computer Science & Engineering, Washington University, St. Louis, MO 63110 USA
[2]School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA
[3]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545 USA
[4]Department of Electrical and System Engineering, Washington University, St. Louis, MO 63110 USA

CORRESPONDING AUTHOR: MICHAEL T. MCCANN (e-mail: mccann@lanl.gov).

This work was supported by the Los Alamos National Laboratory through Laboratory Directed Research and Development Program under Project 20230771DI.

ABSTRACT Ptychography is an imaging technique that captures multiple overlapping snapshots of a sample, illuminated coherently by a moving localized probe. The image recovery from ptychographic data is generally achieved via an iterative algorithm that solves a nonlinear phase retrieval problem derived from measured diffraction patterns. However, these iterative approaches have high computational cost. In this paper, we introduce PtychoDV, a novel deep model-based network designed for efficient, high-quality ptychographic image reconstruction. PtychoDV comprises a vision transformer that generates an initial image from the set of raw measurements, taking into consideration their mutual correlations. This is followed by a deep unrolling network that refines the initial image using learnable convolutional priors and the ptychography measurement model. Experimental results on simulated data demonstrate that PtychoDV is capable of outperforming existing deep learning methods for this problem, and significantly reduces computational cost compared to iterative methodologies, while maintaining competitive performance.

INDEX TERMS Ptychography, deep unrolling, vision transformer, deep learning, and image reconstruction.

I. INTRODUCTION

Ptychography is an essential imaging technique applied in fields such as materials science, biology, and nanotechnology, due to its ability to provide high-resolution images of samples [1]. In ptychographic imaging, a localized coherent scanning probe is moved across a sample while recording a set of far-field diffraction patterns by measuring the intensity of the diffracted waves. The probe is positioned such that each illuminated area has considerable overlap with neighboring regions, providing redundant information that can be used to computationally retrieve the relative phase of recorded intensity data within the Fraunhofer diffraction plane. An estimate of the complex image representing the refractive index and thickness of the object can be obtained from the ptychographic measurements by solving a phase-retrieval optimization problem. A variety of iterative algorithms have been proposed to solve this problem, the main concepts including batch improvement [2], [3], [4] and stochastic or preconditioned gradient approaches [5], [6], [7], [8], [9]. Although these methods have demonstrated satisfactory performance, they suffer from high computational cost due to their iterative nature.

Deep learning (DL) has attracted attention for ptychography due to its potential to reduce the computational cost of ptychographic image reconstruction [10], [11]. Existing techniques depend on *convolutional neural network (CNN)* architectures that directly map measurements to ground truth image patches. Despite being faster than iterative alternatives, CNN-based methods have yet to deliver results comparable with those of iterative methods. This is presumably because exiting CNNs process individual ptychographic measurements in isolation, thereby preventing the exploitation of the *ptychographic measurement model*, such as the redundant information from overlapping illuminated regions. On the other hand, *deep model-based architectures (DMBA)* have shown improved performance over generic CNNs by exploiting the measurement model of imaging problems [12], [13], [14], [15]. A widely-used example of DMBA is the *deep unrolling network (DU)* that interprets iterative

algorithms as a neural network by stacking iterations into layers and then training it end-to-end. Although DU has shown promising results in many imaging problems, to the best of our knowledge, its potential in the context of ptychographic image reconstruction remains unexplored.

In this paper, we bridge this gap by proposing a novel **d***eep unrolling network for* **ptycho***graphic image reconstruction based on* **v***ision transformer (PtychoDV)* that leverages the measurement model to improve DL performance while maintaining low computational cost. Our key contributions in this work are summarized as follows:

- PtychoDV consists of a *vision transformer (ViT)* [16] followed by a DU network. ViT employs self-attention mechanisms that learn the interdependencies between measurements and then reconstructs the entire set of data, providing an initial image for DU. This is essential due to the non-convex and nonlinear nature of ptychography, which makes it nontrivial to direct estimation of an initial image from raw data. DU then refines the initial image by alternating between imposing CNN priors and applying the update rule of Wirtinger flow [8] based on the measurement model.
- We tested PtychoDV on simulated data, demonstrating that it (a) achieved state-of-the-art performance compared with DL baselines, (b) obtained competitive results compared with iterative approaches, with substantially reduced computational cost, and (c) has potential for the sparse sampling setup and providing a suitable initialization for iterative methods, even when the probe in testing differs from that in training.

II. RELATED WORK

In this section, we introduce the notions and related works required to define PtychoDV. We also discuss iterative algorithms and deep learning approaches for ptychographic image reconstruction.

A. PROBLEM FORMULATION

Ptychographic image reconstruction is usually formulated as an inverse problem that recovers an unknown image $\boldsymbol{x} \in \mathbb{C}^n$ from a set of measurements $\{\boldsymbol{y}_i \in \mathbb{R}^m\}_i^N$ characterized by nonlinear systems

$$\boldsymbol{y}_i^2 \sim \mathsf{Pois}(|\boldsymbol{F}\boldsymbol{P}\boldsymbol{D}_i\boldsymbol{x}|^2)\,, \tag{1}$$

where $\boldsymbol{P} \in \mathbb{C}^{m\times m}$ is the complex probe illumination, $\boldsymbol{F} \in \mathbb{C}^{m\times m}$ represents the Fourier transform, $\mathsf{Pois}(\cdot)$ denotes a Poisson distribution that models the detector response, and $|\cdot|$ is an elementwise absolute value operator. In this study, we assume the probe is known and only estimates the image. In (1), $\boldsymbol{D}_i \in \{0,1\}^{m\times n}$ indicates an operator that extracts one patch from $\boldsymbol{x}$, determined by the ith probe location during imaging, and N is the total number of probe locations. Note that we do not consider subpixel illumination shifts For ease of notation in our discussion, we also define $\boldsymbol{x}_i = \boldsymbol{D}_i\boldsymbol{x}$ as a patch of ground truth corresponding to the ith probe location, $\boldsymbol{D}_i^{\mathsf{T}} \in \mathbb{C}^{n\times m}$ as the adjoint operator of $\boldsymbol{D}_i$ that transforms a patch into an image by zero-filling the surplus regions. A common way to solve this inverse problem is to formulate it as an optimization problem

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}} \left\{ \sum_{i=1}^{N} f_i(\boldsymbol{x}_i) \right\}, \tag{2}$$

where

$$f_i(\boldsymbol{x}_i) = \frac{1}{2\sigma_i^2} \left\| \boldsymbol{y}_i - |\boldsymbol{F}\boldsymbol{P}\boldsymbol{x}_i| \right\|^2, \tag{3}$$

represents a cost function enforcing data consistency between $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$. This choice of cost function can be derived as an approximation of the maximum likelihood (ML) cost function for a Poisson noise model [9].

B. ITERATIVE METHODS

A variety of numerical iterative algorithms have been proposed for solving (2) [2], [3], [4], [5], [6], [7], [8], [9]. Many of these methods concurrently update the image patches $\{\widehat{\boldsymbol{x}}_i\}$ and the combine these patches into an estimate image [2], [3], [4], aiming to overcome the computational challenges posed by the substantial volume of data. For example, SHARP [2] relies on alternating projections between constraints in the Fourier domain and image domain. *projected multi-agent consensus equilibrium (PMACE)* [3], [4] solves ptychography problem (2) by finding an equilibrium point $\boldsymbol{x}^*$ that satisfies the equation $[F_1(\boldsymbol{x}_1), \ldots, F_N(\boldsymbol{x}_N)]^T = [\bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_N]^T$, where

$$F_i(\boldsymbol{x}_i) = \arg\min_{\boldsymbol{v}} \left\{ f_i(\boldsymbol{v}) + \frac{1}{2\sigma^2} \|\boldsymbol{F}\boldsymbol{P}\boldsymbol{v} - \boldsymbol{F}\boldsymbol{P}\boldsymbol{x}_i\| \right\}, \tag{4}$$

is derived as a proximal map for $f_i(\boldsymbol{x}_i)$, and

$$\bar{\boldsymbol{x}}_i = \boldsymbol{D}_i\boldsymbol{\Lambda}^{-1} \sum_{i=1}^{N} \boldsymbol{D}_i^{\mathsf{T}} |\boldsymbol{P}|^{\kappa} \boldsymbol{x}_i\,, \tag{5}$$

appropriately averages the estimated patches associated with the same scan locations. In (5), $\boldsymbol{\Lambda} = \sum_{i=1}^{N} \boldsymbol{D}_i^{\mathsf{T}} |\boldsymbol{P}|^{\kappa}$, and κ denotes a probe exponent parameter. Another class of algorithms use stochastic or preconditioned gradient methods to directly refine the estimated image [5], [6], [7], [8], [9]. For instance, *Wirtinger flow (WF)* [8] and *accelerated WF (AWF)* [9] use gradient descent to minimize the non-differentiable objective in (2) by defining a generalized gradient based on the notion of Wirtinger derivatives (see also Sec. VI in [8]). While these methods can provide satisfactory performance, they suffer from high computational cost due to the iterative refinement nature.

C. DEEP LEARNING APPROACHES

Deep learning has gained popularity in the broader context of imaging inverse problems due to its excellent performance (see recent reviews in [14], [17], [18]). A widely-used DL approach is to train a CNN to learn a mapping from the measurements to the desired reconstruction [19], [20]. Several DL methods based on CNNs have been proposed for ptychographic image reconstruction [10], [11], [21], [22]. PtychoNet [10] and PtychoNN [11] involve training an end-to-end DL model by sequentially mapping measurements $\boldsymbol{y}_i$ to corresponding ground truth $\boldsymbol{x}_i$. In testing, one can derive reconstructed images using the raw measurements as inputs to the pre-trained model. These methods can achieve fast reconstruction, but at the expense of performance. We posit that this is due to CNNs processing individual ptychographic measurements, which fundamentally prevents them from exploiting information from the measurement model, such as the redundancy from the overlapping measured diffraction patterns. In this study, we propose to tackle these issues by leveraging two recent approaches: vision transformer (ViT) and deep model-based architecture (DMBAs), detailed discussions of which follow.

ViTs represents a significant shift in computer vision, moving from convolutional architectures to a transformer-based approach (see e.g. recent reviews [23], [24]). The central concept behind ViT is treating image patches as data sequences, and then employing *self-attention mechanisms* to compute attention scores among all

TABLE 1. Quantitative Evaluation of Several Methods With Format of $A \pm B(c)$ on Testing Noisy Measurements, Where A, B and c Denote Mean of Normalized Root Mean-Square-Error (NRMSE), Standard Deviation of NRMSE, and Testing Time (Seconds Per Image), Respectively.

Sampling pattern	256:5	121:8	64:11	25:19	16:27
PtychoNet [10]	0.483 ± 0.56 (0.175)	0.483 ± 0.56 (0.075)	0.483 ± 0.56 (0.042)	0.483 ± 0.56 (0.017)	0.484 ± 0.56 (0.012)
Unet [41]	0.465 ± 0.55 (0.366)	0.465 ± 0.55 (0.165)	0.465 ± 0.55 (0.084)	0.466 ± 0.55 (0.034)	0.467 ± 0.55 (0.022)
ViT [16]	0.441 ± 0.59 (0.062)	0.442 ± 0.59 (0.030)	0.443 ± 0.59 (0.020)	0.447 ± 0.59 (0.009)	0.450 ± 0.59 (0.009)
AWF [9]	0.047 ± 0.15 (109.60)	0.054 ± 0.20 (51.29)	0.071 ± 0.21 (26.12)	0.118 ± 0.34 (10.70)	0.201 ± 0.63 (7.15)
PMACE [4]	**0.035 ± 0.18 (138.08)**	**0.044 ± 0.13 (66.43)**	**0.065 ± 0.19 (34.94)**	0.119 ± 0.33 (13.82)	0.184 ± 0.52 (8.78)
ViT+Unet	0.219 ± 0.65 (0.059)	0.241 ± 0.68 (0.025)	0.254 ± 0.64 (0.017)	0.284 ± 0.65 (0.010)	0.307 ± 0.67 (0.011)
ViT+GD	0.259 ± 0.38 (0.177)	0.261 ± 0.38 (0.081)	0.281 ± 0.41 (0.045)	0.897 ± 0.30 (0.021)	0.929 ± 0.32 (0.015)
ViT+1DU	0.128 ± 0.54 (0.118)	0.139 ± 0.56 (0.051)	0.164 ± 0.56 (0.031)	0.218 ± 0.64 (0.016)	0.245 ± 0.63 (0.015)
Initializer+DU	0.069 ± 0.25 (0.362)	0.076 ± 0.28 (0.183)	0.093 ± 0.34 (0.122)	0.137 ± 0.41 (0.085)	0.173 ± 0.49 (0.077)
PtychoNet+DU	0.046 ± 0.19 (0.617)	0.053 ± 0.22 (0.344)	0.066 ± 0.30 (0.265)	0.104 ± 0.37 (0.245)	0.139 ± 0.46 (0.239)
PtychoDV	0.043 ± 0.19 (0.212)	0.050 ± 0.23 (0.109)	**0.065 ± 0.32 (0.074)**	**0.098 ± 0.36 (0.049)**	**0.127 ± 0.45 (0.044)**

The results with the best and second best mean NRMSE are highlighted. This table shows that PtychoDV can outperform existing DL baseline methods. This table also demonstrates that PtychoDV can gain competitive performance compared against state-of-the-art iterative algorithm, while maintaining significantly lower computational cost.

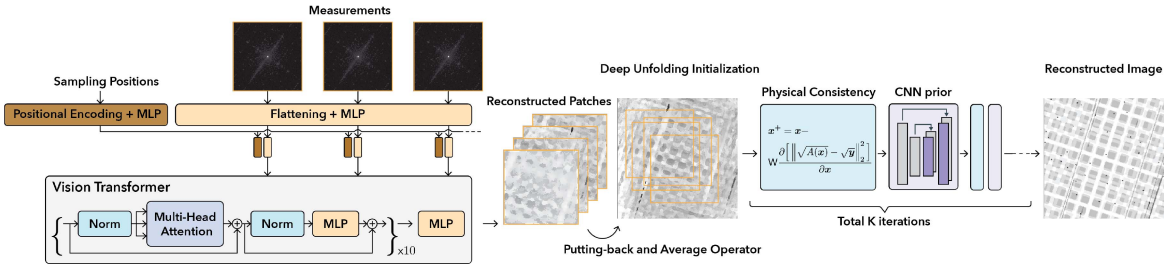

FIGURE 1. Illustration of the pipeline of PtychoDV that consists of two main components: (a) a vision transformer module that reconstructs an initial image from raw measurements by taking into account the interdependencies of the measurements, and (b) a DU network that refines the initial image using the measurement forwards and CNN priors. See (8) for the iterative update of the physical consistency module.

patch pairs, gauging their reciprocal influence. This approach allows each patch to consider all others in its context, efficiently capturing long-range dependencies and complex interrelationships, irrespective of spatial distance. Recent studies have applied ViT in many imaging inverse problems (see Sec. 3.6 in [24]). In ptychography, it is straightforward to apply ViT by considering measurements as a sequence so that their interdependencies can be learned. Despite that, our empirical results in Tables 1 and 3 show that, while ViT can perform better than CNNs, the performance of ViT is inferior to that of iterative approaches. A recent abstract investigated the use of transformers for ptychography [25]. Nonetheless, our work distinguishes itself from [25] in two key aspects: *(a)* our analysis of the algorithm and numerical validation is more extensive, and *(b)* we improve ViT by integrating DU into our proposed pipeline.

DMBAs represent a family of DL algorithms that sconnect measurement models and deep neural networks for solving imaging inverse problems (see also reviews in [14], [15]). Examples of DMBAs include *plug-and-play (PnP)* [12], [26], *regularization by denoiser (RED)* [13], *deep unrolling (DU)* [27], [28], [29], [30], [31], [32], [33], [34], and *deep equilibrium models (DEQ)* [35], [36], [37]. PnP and RED represent classes of iterative algorithms that leverage pre-trained denoisers as imaging priors. A recent study has extended this idea to ptychographic image reconstruction [38]. However, its iterative nature inherently results in a high computational cost. DU has recently gained significant popularity due to its excellent performance and low computational cost. The key idea of DU is to *(a)* implement a finite number of iterations of an image reconstruction algorithm as layers of a network, *(b)* represent the regularization within the iterative algorithm as a trainable CNN, and *(c)* train the resulting network end-to-end. Many recent studies have shown the potential of DU in various imaging inverse problems, including compressed sensing MRI [27], [28], [29], [30], sparse view CT [31], [32], [33], and phase retrieval [34]. A recent study has explored DU in the context of ptychography [39], but it lacks a trainable network for providing dedicated initial images. Different DU architectures can be obtained by using different iterative algorithms. As will be discussed in the next section, our main contribution is to propose a deep unrolling network based on the WF algorithm to significantly improve the deep learning method performance in ptychographic image reconstruction.

III. PROPOSED METHOD: PTYCHODV

As illustrated in Fig. 1, PtychoDV consists of two neural networks: (a) a vision transformer g_θ that estimates initial results from the raw measurements, and (b) a DU network that iteratively refines the initial results. We rely on supervised learning to jointly optimize these two neural networks.

A. VISION TRANSFORMER

The vision transformer g_θ in PtychoDV takes as input a set of raw measurements $\boldsymbol{y}_i$ and reconstructs image patches $\hat{\boldsymbol{x}}_i = \mathsf{g}_\theta(\boldsymbol{y}_i)$. Specifically, the raw measurements are transformed into measurement latent vectors in parallel by a multi-layer perceptron (MLP). The Cartesian coordinates of the corresponding sampling position $\boldsymbol{c}$ are mapped to coordinate latent vectors with the same dimension as the

measurement latent vectors using Fourier positional encoding [40] followed by a MLP

$$\mathsf{MLP}\left(\sin(2^0\pi\boldsymbol{c}), \cos\left(2^0\pi\boldsymbol{c}\right)\ldots\sin\left(\underbrace{2^{L_f}\pi\,\boldsymbol{c}}_{k_{\sin}}\right), \cos\left(\underbrace{2^{L_f}\pi\,\boldsymbol{c}}_{k_{\cos}}\right)\right), \tag{6}$$

where $\sin(\cdot)$ and $\cos(\cdot)$ are element-wise operators. The measurement feature vectors and the coordinate feature vectors are concatenated and then iteratively processed by attention modules, which consist of layer normalization, multi-head self-attention (MHSA) modules, and MLPs. The output feature vectors from the last attention module are transformed into reconstructed patches with the same dimensions as the raw measurements using a output MLP. Further technical details of ViT can be found in [16]. The key differences compared to the original VIT [16] include a different positional embedding derived from the sampling position of ptychography and a modified output layer that transforms the final feature maps of the transformer into patches with dimensions matching those of the raw measurements.

The main innovation behind the use of ViT in g_θ is considering the measurements related to the same ground truth as a sequence. The motivation behind this is to allow the model to capture long-range dependencies and complex relationships among the measurement patches, especially those that overlap, reflecting the imaging nature of ptychography. On the other hand, existing DL methods, such as PtychoNet [10], reconstruct the measurements in parallel, without taking into account dependencies among measurements.

We then convert the reconstructed patches into an image $\widehat{\boldsymbol{x}}$ by the following steps: *(a)* we initialize $\widehat{\boldsymbol{x}}$ as an all-zero image and create counters for each pixel location; *(b)* we add each reconstructed patch to the corresponding sampling region in $\widehat{\boldsymbol{x}}$ and increase the counters in that area by one; and *(c)* We perform element-wise division of $\widehat{\boldsymbol{x}}$ by the counter at all locations where the counter has non-zero value.

B. DEEP UNROLLING NETWORK

The DU network in PtychoDV is obtained by interrupting the iteration of the proximal gradient PnP framework [26] which consists of K iterations of gradient descent each followed by neural network refinement

$$\widehat{\boldsymbol{x}}^{k+1} = \mathsf{h}_{\boldsymbol{\varphi}}\left(\widehat{\boldsymbol{x}}^k - \mathsf{WF}(\widehat{\boldsymbol{x}}^k)\right)\ \forall k = 0,\ldots K-1\,, \tag{7}$$

where $\mathsf{h}_{\boldsymbol{\varphi}}$ denotes a CNN with trainable parameter $\boldsymbol{\varphi}\in\mathbb{R}^m$, $\widehat{\boldsymbol{x}}^{k+1}$ is the output of the kth layer of DU, and $\widehat{\boldsymbol{x}}^0 = \widehat{\boldsymbol{x}}$. Here, $\mathsf{WF}(\cdot)$ represents a Wirtinger flow gradient update of the objective in (2)

$$\mathsf{WF}\left(\widehat{\boldsymbol{x}}^k\right) = \gamma\sum_i^N \boldsymbol{D}_i^{\mathsf{T}}\boldsymbol{P}^{\mathsf{H}}\boldsymbol{F}^{\mathsf{H}}\left(\boldsymbol{m}_i\left(\widehat{\boldsymbol{x}}^k\right) - \boldsymbol{y}_i\frac{\boldsymbol{m}_i(\widehat{\boldsymbol{x}}^k)}{|\boldsymbol{m}_i(\widehat{\boldsymbol{x}}^k)|}\right), \tag{8}$$

where

$$\boldsymbol{m}_i(\widehat{\boldsymbol{x}}^k) = \boldsymbol{F}\boldsymbol{P}\boldsymbol{D}_i\widehat{\boldsymbol{x}}^k\,, \tag{9}$$

$(\cdot)^{\mathsf{H}}$ is the conjugate transpose, and γ represents a step size of $\max(\sum_{i=1}^N \boldsymbol{D}_i^{\mathsf{T}}|\boldsymbol{P}|^2)$. The WF gradient descent allows DU to exploit the information from the physical model of ptychography by fitting the intermediate estimation to the raw measured data. $\mathsf{h}_{\boldsymbol{\varphi}}$ further refines the estimation by imposing a prior information learned from the external dataset.

C. LOSS FUNCTION

We trained g_θ and $\mathsf{h}_{\boldsymbol{\varphi}}$ jointly in an end-to-end manner by minimizing the loss function

$$\ell_{\mathsf{loss}} = \ell_{\mathsf{image}}(\boldsymbol{\varphi}) + \lambda\,\ell_{\mathsf{patch}}(\boldsymbol{\theta})\,, \tag{10}$$

where λ is a trade-off parameter. The purpose of (10) is to promote high-quality reconstruction in both the image-wise and patch-wise manners. Specifically, ℓ_{image} is formulated to penalize the difference between the final estimation of DU and the corresponding ground truth

$$\ell_{\mathsf{image}} = \left\|\widehat{\boldsymbol{x}}^K - \boldsymbol{x}\right\|^2\,, \tag{11}$$

and ℓ_{patch} seeks to minimize the discrepancy between estimated patches of ViT and the corresponding ground truth patches

$$\ell_{\mathsf{patch}} = \sum_{i=1}^N \|\widehat{\boldsymbol{x}}_i - \boldsymbol{x}_i\|^2\,. \tag{12}$$

IV. NUMERICAL VALIDATION

This section presents the setup and results of our numerical validation on PtychoDV. We discuss our dataset, the implementation of PtychoDV, our comparison method, and our evaluation metrics.

A. EXPERIMENTAL SETUP

1) DATASET

We simulated a dataset consisting of ground truth complex-valued images (*i.e.,* $\boldsymbol{x}$ in (1)) and ground truth complex-valued probes (*i.e.,* $\boldsymbol{P}$ in (1)). Ground truth images were 400×400 pixels and had assigned density and thickness to model a multi-layer Copper-Tungsten composite material. Simulated probes were 256×256 pixels with a photon energy of 8.8 keV. We simulated 60,000, 100, and 100 ground truth samples for training, validation, and testing, respectively. We simulated two types of probes, which we shall refer to as *probe A* and *probe B*. We used *probe A* to generate datasets for training and testing, while *probe B* was used only for testing, in order to evaluate the generalization of the pre-trained model on measurements simulated using an unseen probe. *Probe B* was assumed to be unknown in this experiment, while *probe A* was known. Different sampling patterns (i.e., $\boldsymbol{D}_i$ in (1)) were simulated, denoted as N:L, where the probe locations form an $\sqrt{N}\times\sqrt{N}$ grid with grid spacing equal to L pixels. We experimented with N:L values of 256:5, 121:8, 64:11, 25:19, and 16:27. The smaller the value of N, the sparser the sampling pattern. The training dataset involves different sampling patterns. Fig. 2 illustrates a sample of ground truth images, 256:5 sampling patterns, and the simulated ground truth probes. We followed [4] to use r_p, the peak photon rate, to scale the mean of a Poisson distribution to obtain noisy simulated measurements

$$\hat{\boldsymbol{y}}_i^2 \sim \mathsf{Pois}\left(\frac{|\boldsymbol{F}\boldsymbol{P}\boldsymbol{x}_i|^2}{\max(|\boldsymbol{F}\boldsymbol{P}\boldsymbol{x}_i|^2)}\times r_p\right)\,. \tag{13}$$

As r_p increases, the signal-to-noise ratio also increases. Assuming a photon detector with 14-bit dynamic range, we take $r_p = 10^5$ for our simulated noisy diffraction patterns. Fig. 2 illustrates a sample of ground truth images, two stimulated ground truth probes, and sampling pattern of 256:5.

2) IMPLEMENTATION

We experimented with several values of λ in (10). The best empirical results were obtained when $\lambda = 1$. We set the number of DU iteration

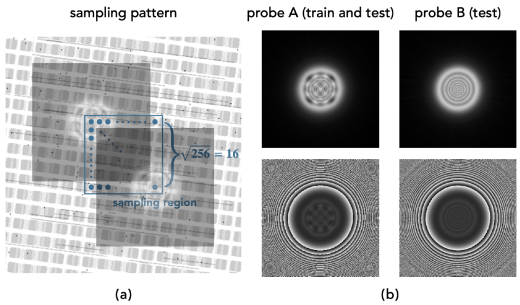

FIGURE 2. (a) Magnitude of a ground truth image illustrating a sampling pattern of 256:5. (b) Two simulated ground truth probes (top row: magnitude, bottom row: phase). Probe *A* was used to simulate measurements for training and testing, while probe *B* was exclusively for testing the pre-trained models.

of PtychoDV to $K = 3$, which is the maximum number achievable under the memory constraints of our workstation. We set L_f in (6) to 10. We used the Adam [42] optimizer with learning rate 10^{-5} and mini-batch size 1, training for 30 epochs. We performed all experiments on a host equipped with an AMD Ryzen Threadripper 3960X Processor and an NVIDIA GeForce RTX 3090 GPU. The training time of PtychoDV on this host was around 120 hours. Our PtychoDV implementation is publicly available[1].

3) EVALUATION

We followed [4] in using normalized root-mean-square error (NRMSE) to evaluate the quality of reconstructed images. Because that the measured data is not sensitive to a constant phase shift in the full transmittance image, we have taken into account this phase shift while calculating the NRMSE between the reconstructed complex image $\widehat{\boldsymbol{x}}$ and the ground truth image $\boldsymbol{x}$. Specifically, the NRMSE is calculated elementwise as follows:

$$\text{NRMSE}(\hat{x}, x) = \frac{|\hat{x} - e^{i\theta}x|}{|x|} \tag{14}$$

where $\theta \in [0, 2\pi)$ is chosen to minimize the numerator.

4) COMPARISON

We compared PtychoDV with several baseline approaches, including PtychoNet [10], Unet, ViT, PMACE [4] and AWF [9]. PtychoNet is a DL method that uses a CNN to map individual measurements directly to the corresponding ground truth image patch. We implemented PtychoNet, Unet, and ViT. For PMACE and AWF, we used the official implementations from the PMACE repository[2]. We set the total number of iterations of PMACE and AWF to 100. We followed [4] to estimate the initial images for PMACE and AWF. Unet and ViT is similar to PtychoNet, but having more complex neural network architectures.

In order to determine the impact of different elements in our configuration, we conducted a component analysis with various versions of PtychoDV, termed as *ViT+Unet*, *ViT+GD*, *ViT+1DU*, *Initializer+DU* and *PtychoNet+DU*. *ViT+Unet* replaces the DU with Unet, thereby removing the integration of the measurement models in the resulting network architecture. *ViT+GD* excludes the CNN priors in DU, whereas *ViT+1DU* reduces the number of DU iterations to one. *PtychoNet+DU* substitutes ViT with PtychoNet as the CNN used for computing the initial images. *Initializer+DU* substitutes ViT with a handcrafted initialization approach (refer to equation (24) in [4]). The trainable components of *Initializer+DU* constitute a pure deep unrolling architecture.

In addition, we tested the use of the PtychoDV reconstructions as initialization for PMACE. We conducted experiments on both *probe A* and *probe B*. The resulting methods are as follows: *(a) PtychoDV-A* tests PtychoDV on testing data stimulated using *probe A*; *(b) PMACE-A* tests PMACE on testing data stimulated using *probe A*; *(c) PMACE-A-10* is a variant of *PMACE-A* with total number of iterations being 10; *(d) PMACE-A-10 w/ PtychoDV* is similar to *PMACE-A-10* but use PtychoDV to estimate the initial image; *(e) PtychoDV-B* tests PtychoDV on testing data stimulated using *probe B*; *(f) PMACE-B* tests PMACE on testing data stimulated using *probe B*; *(g) PMACE-B w/ PtychoDV* is similar to *PMACE-B* but use PtychoDV to estimate the initial image. Since *probe B* was assumed to be unknown, PMACE in (f) and (g) was implemented to jointly estimate the image and the probe. We used probe A as the initial probe when jointly estimating probe B.

B. RESULTS

Table 1 provides a quantitative evaluation and testing time for PtychoDV, baseline approaches, and methods with different components during testing of noisy cases with all sampling patterns. As displayed in Table 1, ViT can achieve lower average NRMSE values than Unet and PtychoNet, both of which are CNN-based, but it still performs suboptimally when compared to iterative methods. When comparing PtychoDV with ViT+Unet and ViT+GD, it is evident from Table 1 that DU and h_φ are essential components of PtychoDV for achieving superior imaging quality. The results from ViT+1DU indicate the potential for improving PtychoDV by increasing the number of DU iterations. Table 1 shows that, when comparing with *Initializer+DU* and *PtychoNet+DU*, PtychoDV can gain superior performnace, highlighting the importance of using ViT to compute the initial image. While both ViT and DU serve as necessary components within PtychoDV, Table 1 indicates that incorporating DU leads to higher SNR improvements than incorporating ViT. To conlude, Table 1 demonstrates that PtychoDV can achieve performance that is competitive with, and even superior to (in the sparse sampling pattern), PMACE, the state-of-the-art iterative method. Finally, while PtychoDV is the most time-consuming method among DL baselines, it still has significantly less computational cost than iterative methods across all sampling patterns.

Fig. 3 provides visual results of PtychoDV and baseline methods on noisy cases with *sparse* 64:11 sampling patterns. Fig. 3 shows that end-to-end neural networks, which directly map measurements to ground truth image patches, tend to reconstruct images with blurry details, whereas PtychoDV provides less noisy and sharper images. This figure also highlights that AWF and PMACE, the two commonly used iterative algorithms, reconstruct noisy images with a higher NRMSE than PtychoDV in the sparse 64:11 sampling pattern. Fig. 4 provides visual results of PtychoDV compared to ablated methods on testing noisy cases with the sampling patterns of 64:11. This figure illustrates that PtychoDV can quantitatively and qualitatively outperform several ablated variants.

Table 2 provide a quantitative evaluation and testing time for PtychoDV and PMACE on noisy testing data, simulated with different probes and different sampling patterns. These tables demonstrate

[1] https://github.com/wjgancn/PtychoDV
[2] https://github.com/cabouman/ptycho_pmace

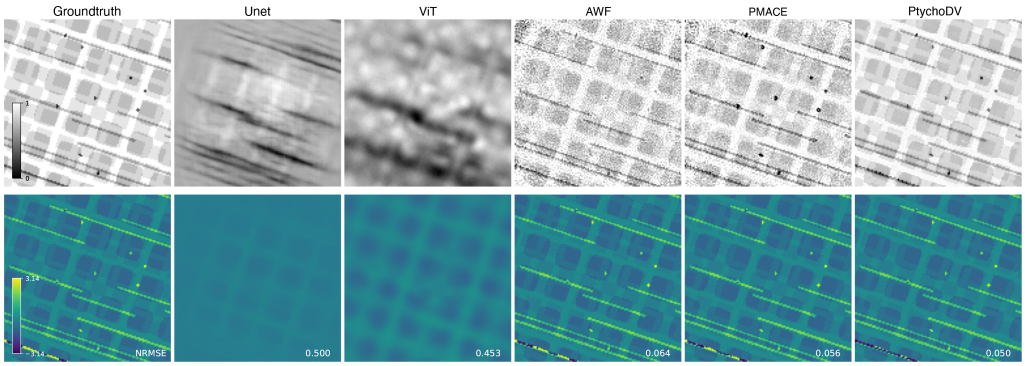

FIGURE 3. Visual results of PtychoDV and other baseline methods on noisy testing data with sampling pattern of 64:11. The magnitude and the phase of the reconstructed images are shown in the top and the bottom row, respectively. NRMSE values are included in the right bottom of each image. This figure highlights superior performance of PtychoDV on sparse sampling pattern. Note that PtychoDV can reconstruct images that are consistent with ground truth, whereas the results from the other baseline exhibit noise and blurry artifacts.

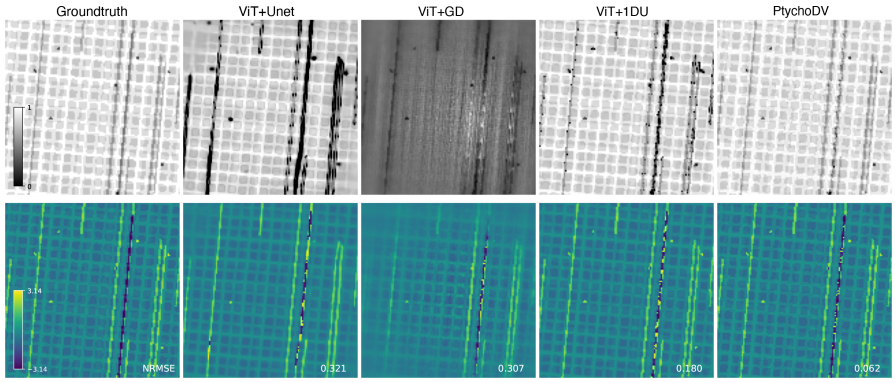

FIGURE 4. Visual results of PtychoDV and its variants on noisy testing data with sampling pattern of 64:11. The magnitude and the phase of the reconstructed images are shown in the top and the bottom row, respectively. NRMSE values of each method is labeled in the right bottom of each image. This figure shows that PtychoDV can gain superior performance over its ablated methods.

TABLE 2. Quantitative Evaluation of Several Methods With Format of $A \pm B(c)$ on Testing Noisy Measurements, Where A, B and c Denote Mean of Normalized Root Mean-Square-Error (NRMSE), Standard Deviation of NRMSE, and Testing Time (Seconds Per Image), Respectively.

Sampling pattern	256:5	121:8	64:11	25:19	16:27
PtychoDV-A	0.044 ± 0.17 (0.212)	0.053 ± 0.23 (0.109)	0.066 ± 0.27 (0.074)	0.102 ± 0.40 (0.049)	0.135 ± 0.49 (0.044)
PMACE-A	**0.035 ± 0.17 (138.08)**	0.045 ± 0.18 (66.43)	0.065 ± 0.20 (34.94)	0.118 ± 0.36 (13.82)	0.193 ± 0.58 (8.78)
PMACE-A-10	0.246 ± 0.61 (16.22)	0.251 ± 0.61 (8.16)	0.264 ± 0.62 (4.62)	0.297 ± 0.65 (2.29)	0.351 ± 0.74 (1.72)
PMACE-A-10 w/ PtychoDV	0.039 ± 0.09 (16.43)	**0.042 ± 0.08 (8.43)**	**0.048 ± 0.08 (5.10)**	**0.066 ± 0.07 (2.34)**	**0.084 ± 0.15 (1.79)**
PtychoDV-B	0.151 ± 0.57 (0.212)	0.157 ± 0.55 (0.109)	0.176 ± 0.65 (0.074)	0.209 ± 0.59 (0.049)	0.245 ± 0.75 (0.044)
PMACE-B	0.332 ± 0.81 (288.91)	0.329 ± 0.84 (137.25)	0.340 ± 0.82 (73.12)	0.412 ± 0.71 (28.51)	0.438 ± 0.64 (17.84)
PMACE-B w/ PtychoDV	**0.081 ± 0.51 (291.54)**	**0.089 ± 0.34 (139.49)**	**0.096 ± 0.53 (75.76)**	**0.139 ± 0.42 (28.93)**	**0.166 ± 0.64 (18.13)**

The results with the best and second best mean NRMSE over the same testing data are highlighted. This table highlights that PtychoDV initialization could significantly reduce number of iteration of PMACE, thus reducing the computational cost, without sacrificing the performance. This table also shows that PtychoDV could also provide good initialization for better imaging quality of PMACE even when the testing probe differs to that used for training.

that, when tested on a *known* probe A, PMACE initialized by PtychoDV can achieve performance competitive with generic PMACE, but with significantly fewer iterations and lower computational cost. The tables also indicate that, when tested on an *unknown* probe B, PMACE initialized by PtychoDV can achieve superior performance compared to PMACE on the joint estimation of image and probe.

V. DISCUSSION AND CONCLUSION

This paper presents PtychoDV, a new DL method for ptychographic image reconstruction. The key idea behind PtychoDV is a deep unrolling architecture that systematically integrates trainable neural network priors and measurement operators of the ptychography. Moreover, we employ a vision transformer to estimate initial images from raw measurements, which allows capturing long-range dependencies in the data effectively.

The major benefits of PtychoDV include its remarkable performance improvements, both quantitatively and qualitatively, compared to existing deep learning methods. Furthermore, PtychoDV achieves competitive performance against existing iterative algorithms, but with a substantially lower computational cost. Moreover, in sparse sampling setup, PtychoDV outperforms iterative methods.

TABLE 3. Quantitative Evaluation of Several Methods With Format of $A \pm B(c)$ on Testing *noise-Free* Measurements, Where *A*, *B* and *c* Denote Mean of Normalized Root Mean-Square-Error (NRMSE), Standard Deviation of NRMSE, and Testing Time (Seconds Per Image), Respectively

Sampling pattern	256:5	121:8	64:11	25:19	16:27
PtychoNet [10]	0.488 ± 0.55 (0.174)	0.488 ± 0.55 (0.075)	0.488 ± 0.55 (0.040)	0.488 ± 0.55 (0.017)	0.489 ± 0.55 (0.012)
Unet [41]	0.455 ± 0.58 (0.362)	0.455 ± 0.58 (0.164)	0.455 ± 0.58 (0.084)	0.456 ± 0.58 (0.034)	0.456 ± 0.58 (0.022)
ViT [16]	0.410 ± 0.59 (0.061)	0.411 ± 0.59 (0.030)	0.413 ± 0.59 (0.018)	0.415 ± 0.59 (0.011)	0.416 ± 0.59 (0.008)
AWF [9]	0.012 ± 0.13 (55.32)	0.013 ± 0.22 (26.61)	0.023 ± 0.22 (14.22)	0.048 ± 0.41 (5.77)	0.107 ± 0.75 (3.94)
PMACE [4]	**0.005 ± 0.10 (76.79)**	**0.006 ± 0.06 (36.67)**	**0.010 ± 0.11 (19.80)**	**0.022 ± 0.26 (7.72)**	0.044 ± 0.48 (5.08)
ViT+Unet	0.174 ± 0.61 (0.056)	0.188 ± 0.60 (0.025)	0.209 ± 0.60 (0.018)	0.240 ± 0.60 (0.012)	0.257 ± 0.62 (0.009)
ViT+GD	0.761 ± 0.51 (0.174)	0.772 ± 0.52 (0.080)	0.767 ± 0.49 (0.046)	0.897 ± 0.30 (0.021)	0.929 ± 0.32 (0.015)
ViT+1DU	0.081 ± 0.36 (0.117)	0.094 ± 0.38 (0.049)	0.116 ± 0.47 (0.030)	0.157 ± 0.56 (0.016)	0.178 ± 0.59 (0.013)
Initializer+DU	0.051 ± 0.22 (0.362)	0.052 ± 0.27 (0.183)	0.063 ± 0.34 (0.122)	0.097 ± 0.56 (0.085)	0.131 ± 0.71 (0.077)
PtychoNet+DU	0.034 ± 0.15 (0.617)	0.034 ± 0.17 (0.344)	0.040 ± 0.21 (0.265)	0.059 ± 0.37 (0.245)	0.084 ± 0.53 (0.239)
PtychoDV	0.013 ± 0.10 (0.211)	0.013 ± 0.10 (0.110)	0.017 ± 0.14 (0.074)	0.028 ± 0.23 (0.049)	**0.038 ± 0.28 (0.044)**

The results with the best and second best mean NRMSE are highlighted. As evidenced by the table, PtychoDV surpasses existing deep learning baseline methods. Moreover, it showcases that PtychoDV can achieve performance comparable to state-of-the-art iterative algorithms while maintaining considerably lower computational cost.

TABLE 4. Quantitative Evaluation of Several Methods With Format of $A \pm B(c)$ on Testing *noise-Free* Measurements, Where *A*, *B* and *c* Denote Mean of Normalized Root Mean-Square-Error (NRMSE), Standard Deviation of NRMSE, and Testing Time (Seconds Per Image), Respectively

Sampling pattern	256:5	121:8	64:11	25:19	16:27
PtychoDV-A	0.014 ± 0.09 (0.211)	0.014 ± 0.12 (0.110)	0.018 ± 0.13 (0.074)	0.030 ± 0.27 (0.049)	0.042 ± 0.32 (0.044)
PMACE-A-10	0.181 ± 0.65 (16.22)	0.188 ± 0.66 (8.16)	0.210 ± 0.67 (4.62)	0.252 ± 0.70 (2.29)	0.316 ± 0.79 (1.72)
PMACE-A	0.006 ± 0.10 (76.79)	0.008 ± 0.14 (36.67)	0.012 ± 0.13 (19.80)	0.025 ± 0.30 (7.72)	0.048 ± 0.48 (5.08)
PMACE-A-10 w/ PtychoDV	**0.003 ± 0.02 (16.43)**	**0.003 ± 0.02 (8.43)**	**0.004 ± 0.03 (5.10)**	**0.006 ± 0.04 (2.34)**	**0.010 ± 0.06 (1.79)**
PtychoDV-B	0.109 ± 0.42 (0.212)	0.111 ± 0.49 (0.109)	0.120 ± 0.49 (0.074)	0.147 ± 0.72 (0.049)	0.163 ± 0.65 (0.044)
PMACE-B	0.321 ± 0.80 (288.91)	0.317 ± 0.82 (137.25)	0.326 ± 0.80 (73.12)	0.391 ± 0.76 (28.51)	0.421 ± 0.69 (17.84)
PMACE-B w/ PtychoDV	**0.017 ± 0.16 (291.54)**	**0.021 ± 0.40 (139.49)**	**0.019 ± 0.19 (75.76)**	**0.045 ± 0.50 (28.93)**	**0.045 ± 0.43 (18.13)**

The results with the best and second best mean NRMSE over the same testing data are highlighted. The information provided in this table demonstrates that by utilizing PtychoDV for initializations, the number of iterations required by PMACE can be significantly reduced, thus lowering computational cost without compromising the performance. It also underlines that even when the probe used for testing is different from the one used during training, PtychoDV can still provide beneficial initialization to enhance PMACE's imaging quality.

The results of PtychoDV show its potential for applications that require real-time reconstruction or fast sampling.

Another important application of PtychoDV is to provide a reliable initialization for existing iterative algorithms. This initialization approach leads to a reduction in the total number of iterations without sacrificing performance. Even in cases where the probe is unknown, iterative algorithms can still benefit from PtychoDV's initialization, regardless of whether the testing probe differs from the training one.

A key feature of PtychoDV is its ability to incorporate and exchange information from all measurements patches simultaneously in the reconstruction. This exchange is technically facilitated through the WF update as described in (8) and the attention module in ViT. Note that the effectiveness of this exchange is also contingent on the overlap probe ratio. Experimental results show that a higher overlap ratio (e.g., sampling pattern of $256:5$) leads to improved performance, indicating enhanced exchange efficiency. In contrast, existing approaches without such deliberate exchange (e.g., PtychoNet) achieve similar performance across different sampling patterns.

The experiments in this study were entirely simulation-based, primarily due to the large number of training pairs and high-quality references required for the proposed method. It is impractical to source such a dataset from real-world samples. Our future direction includes testing PtychoDV on real data and training PtychoDV without high-quality ground truth using self-supervised learning [30].

APPENDIX

This appendix reports experimental results for noise-free cases. We synthesized noise-free measurements as $\boldsymbol{y}_i^2 = |\boldsymbol{F}\boldsymbol{P}\boldsymbol{x}_i|^2$. The other experimental setups are identical to those described in Section IV-A.

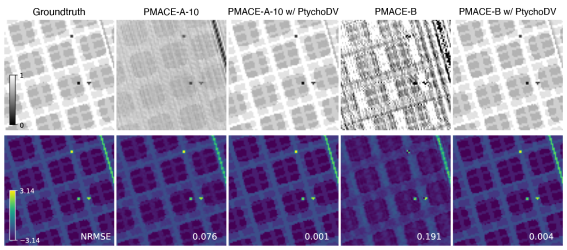

FIGURE 5. Visual results of PMACE tested on noise-free data generated using different probe and different initialization. The magnitude and the phase of the reconstructed images are shown in the top and the bottom row, respectively. NRMSE values of each method is labeled in the right bottom of each image. This figure shows that PMACE with a small number of iterations can achieve better performance by using PtychoDV initialization than that without it. This figure also highlights that PtychoDV could also be used to compute initialization even when the testing probe is different from the probe used in training.

TABLE 5. GPU Memory Usage (GB) for PtychoDV and DL Baseline Methods During Both Training and Inference.

	Inference					Training
Sampling Pattern	256:5	121:8	64:11	25:19	16:27	
PtychoNet [10]	11.84	6.61	4.40	2.88	2.53	21.01
Unet [41]	20.13	10.51	5.68	3.44	2.88	23.12
ViT [16]	4.09	3.71	3.55	3.45	3.41	14.57
ViT+Unet	5.73	4.81	4.47	4.57	4.51	15.46
ViT+GD	4.81	4.08	3.77	3.59	3.51	14.67
ViT+1DU	5.45	4.73	4.54	4.63	4.56	15.47
PtychoNet+DU	11.94	6.71	4.49	2.98	2.64	21.19
Initializer+DU	3.16	2.76	2.71	2.54	2.51	4.22
PtychoDV	5.45	4.73	4.54	4.63	4.56	17.08

TABLE 6. Quantitative Evaluation of PtychoDV and Baseline Methods With Format of $A \pm B$ on a New Noisy Testing Dataset Generated From Probe C, Where A and B Denote Mean of Normalized Root Mean-Square-Error (NRMSE) and Standard Deviation of NRMSE, Respectively.

Sampling pattern	256:5	121:8	64:11	25:19	16:27
Ours-C	0.411 ± 1.16	0.409 ± 1.15	0.413 ± 1.12	0.417 ± 1.05	0.433 ± 1.05
PMACE-C w/o PtychoDV	0.438 ± 1.16	0.423 ± 1.03	0.419 ± 0.77	0.467 ± 0.87	0.483 ± 0.96
PMACE-C w/ PtychoDV	**0.394 ± 1.45**	**0.389 ± 1.40**	**0.389 ± 1.35**	**0.392 ± 1.37**	**0.420 ± 1.11**

The results with the best mean NRMSE are highlighted. Note that probe C is exclusively used for testing. This table highlights that PtychoDV can offer a reliable initialization for PMACE, even when the testing dataset is generated using a more dissimilar asymmetry probe.

TABLE 7. Quantitative Comparison Between the Proposed Loss Function and Its Constituent Parts With Format of $A \pm B$ on Testing Noisy Measurements, Where A and B Denote Mean of Normalized Root Mean-Square-Error (NRMSE) and Standard Deviation of NRMSE, Respectively.

Sampling pattern	256:5	121:8	64:11	25:19	16:27
PtychoDV w/o *image-wise loss*	0.441 ± 0.59	0.442 ± 0.59	0.443 ± 0.59	0.447 ± 0.59	0.450 ± 0.59
PtychoDV w/o *patch-wise loss*	0.047 ± 0.22	0.055 ± 0.26	0.072 ± 0.32	0.111 ± 0.38	0.142 ± 0.43
PtychoDV	**0.043 ± 0.19**	**0.050 ± 0.23**	**0.065 ± 0.32**	**0.098 ± 0.36**	**0.127 ± 0.45**

The results with the best mean NRMSE are highlighted. This table shows that the proposed loss function can gain superior performance over its constituent variants.

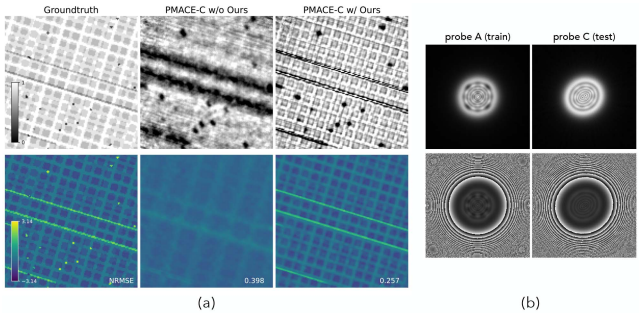

FIGURE 6. (a) Illustration of reconstructed results obtained by PMACE with and without PtychoDV providing initialization. Experiments were conducted on a new testing dataset generated from a new probe C with a sampling pattern of 64 : 11. (b) Illustrations of the training probe A and the new testing probe C. Probe A is used for generating the training dataset, while probe C is exclusively for testing. Note that probe A is symmetrical, whereas probe C is asymmetrical.

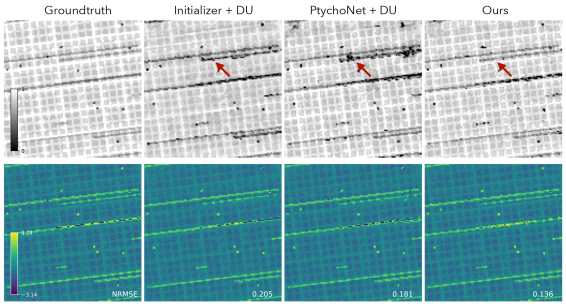

FIGURE 7. Visual results of PtychoDV and other baseline methods on noisy testing data with sampling pattern of 64 : 11. The magnitude and the phase of the reconstructed images are shown in the top and the bottom row, respectively. NRMSE values are included in the right bottom of each image. This figure demonstrates that PtychoDV can provide reconstructions that are more consistent with the ground truth, as highlighted by image features indicated by the red arrow.

Table 3 summarizes the same type of quantitative evaluation and testing time as Table 1, but on noise-free testing data, corroborating the same conclusions drawn from Table 1. Table 4 provide a quantitative evaluation and testing time for PtychoDV and PMACE on noise-free testing data, stimulated with different probes and different sampling patterns. Fig. 5 shows visual results of PMACE on noise-free data with and without initialization generated by PtychoDV. This figure demonstrates that PMACE initialized by PtychoDV can provide results more consistent with the ground truth than those without it.

Table 5 shows memory usage of PtychoDV and other baseline methods, demonstrating that ViT-based methods exhibit lower memory complexity. We attribute this to the smaller dimensions of 1D latent features in ViT compared to the 2D feature maps in CNN.

We also validated PtychoDV on testing dataset generated from an asymmetry *probe C*. This new *probe C* is more dissimilar to *probe A* compared to *probe B*. We tested the application of PtychoDV for providing a reliable initialization for PMACE on this new dataset. Table 6 and Fig. 6 show quantitative and visual results on new testing dataset, respectively. Both Table 6 and Fig. 6 demonstrate that PtychoDV can provide a reliable initialization for PMACE even when the testing dataset is generated using a more dissimilar asymmetry probe.

We performed experiments comparing the proposed loss function and its constituent parts. We summarized the quantitative results in Table 7. Table 7 shows that the proposed loss function can provide superior performance over its constituent variants.

Fig. 7 illustrates the reconstructed images of PtychoNet+DU, Initilizer+DU and PtychoDV. Fig. 7 shows that PtychoDV can provide reconstructions that are more consistent with the ground truth, as highlighted by image features indicated by the red arrow.

ACKNOWLEDGMENT

The authors thank John Barber for his assistance in using the LANL wave propagation code WavePro to generate the training data used in this work.

REFERENCES

[1] F. Pfeiffer, "X-ray ptychography," *Nature Photon*, vol. 12, no. 1, pp. 9–17, Jan. 2018.
[2] S. Marchesini et al., "SHARP: A distributed, GPU-based ptychographic solver," *J. Appl. Crystallogr.*, vol. 49, no. 4, pp. 1245–1252, Aug. 2016.
[3] Q. Zhai, G. T. Buzzard, K. M. Mertes, B. Wohlberg, and C. A. Bouman, "Projected multi-agent consensus equilibrium (PMACE) with application to ptychography," *IEEE Trans. Comput. Imag.*, vol. 9, pp. 1058–1070, 2023.

[4] Q. Zhai, B. Wohlberg, G. T. Buzzard, and C. A. Bouman, "Projected multi-agent consensus equilibrium for ptychographic image reconstruction," in *Proc 55th Asilomar Conf. Signals Syst. Comput.*, 2021, pp. 1694–1698.

[5] J. M. Rodenburg and H. M. L. Faulkner, "A phase retrieval algorithm for shifting illumination," *Appl. Phys. Lett.*, vol. 85, no. 20, pp. 4795–4797, Nov. 2004.

[6] A. M. Maiden and J. M. Rodenburg, "An improved ptychographical phase retrieval algorithm for diffractive imaging," *Ultramicroscopy*, vol. 109, no. 10, pp. 1256–1262, Sep. 2009.

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

[8] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Trans. Inform. Theory*, vol. 61, no. 4, pp. 1985–2007, Apr. 2015.

[9] R. Xu et al., "Accelerated Wirtinger flow: A fast algorithm for ptychography," Jun. 2018, *arXiv:1806.05546*.

[10] Z. Guan, E. Tsai, X. Huang, K. Yager, and H. Qin, "PtychoNet: Fast and high quality phase retrieval for ptychography," Brookhaven Nat. Lab. (BNL), Upton, NY, USA, Tech. Rep. Tech. Rep. BNL-213637-2020-FORE, Sep. 2019.

[11] M. J. Cherukara et al., "AI-enabled high-resolution scanning coherent diffraction imaging," *Appl. Phys. Lett.*, vol. 117, no. 4, Apr. 2020, Art. no. 044103.

[12] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, "Plug-and-play priors for model based reconstruction," in *Proc. IEEE Glob. Conf. Signal Process. Inf. Process*, Dec. 2013, pp. 945–948.

[13] 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, Jan. 2017.

[14] G. Ongie, A. Jalal, C. A. Metzler, R. G. Baraniuk, A. G. Dimakis, and R. Willett, "Deep learning techniques for inverse problems in imaging," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 39–56, May 2020.

[15] 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.

[16] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[17] M. T. McCann, K. H. Jin, and M. Unser, "Convolutional neural networks for inverse problems in imaging: A review," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 85–95, Nov. 2017.

[18] A. Lucas, M. Iliadis, R. Molina, and A. K. Katsaggelos, "Using deep neural networks for inverse problems in imaging: Beyond analytical methods," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 20–36, Jan. 2018.

[19] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4509–4522, Sep. 2017.

[20] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen, "Image reconstruction by domain-transform manifold learning," *Nature*, vol. 555, no. 7697, pp. 487–492, Mar. 2018.

[21] S. Barutcu, D. Gürsoy, and A. K. Katsaggelos, "Compressive ptychography using deep image and generative priors," May 2022, *arXiv:2205.02397*.

[22] F. Guzzi, G. Kourousias, F. Billé, R. Pugliese, A. Gianoncelli, and S. Carrato, "A parameter refinement method for ptychography based on deep learning concepts," *Condens. Matter*, vol. 6, no. 4, Oct. 2021, Art. no. 36.

[23] Y. Liu et al., "A survey of visual transformers," *IEEE Trans. Neural Netw. Learn.*, early access, Mar. 30, 2023, doi: 10.1109/TNNLS.2022.3227717.

[24] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, 2022.

[25] I. Kang et al., "Three-dimensional reconstruction of integrated circuits by single-angle X-ray ptychography with machine learning," in *Proc. Comput. Opt. Sens. Imag.*, 2021, Paper CTu6A.4.

[26] U. S. Kamilov, C. A. Bouman, G. T. Buzzard, and B. Wohlberg, "Plug-and-play methods for integrating physical and learned models in computational imaging," *IEEE Signal Process. Mag.*, vol. 40, no. 1, pp. 85–97, Jan. 2023.

[27] J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, and D. Rueckert, "A deep cascade of convolutional neural networks for dynamic MR image reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 2, pp. 491–503, Feb. 2018.

[28] K. Hammernik et al., "Learning a variational network for reconstruction of accelerated MRI data," *Magn. Reson. Med.*, vol. 79, no. 6, pp. 3055–3071, 2018.

[29] H. K. Aggarwal, M. P. Mani, and M. Jacob, "MoDL: Model-based deep learning architecture for inverse problems," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 394–405, Feb. 2019.

[30] Y. Hu et al., "SPICE: Self-supervised learning for MRI with automatic coil sensitivity estimation," 2022, *arXiv:2210.02584*.

[31] J. Adler and O. Oktem, "Learned primal-dual reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1322–1332, Jun. 2018.

[32] W. Wu, D. Hu, C. Niu, H. Yu, V. Vardhanabhuti, and G. Wang, "DRONE: Dual-domain residual-based optimization network for sparse-view ct reconstruction," *IEEE Trans. Med. Imag.*, vol. 40, no. 11, pp. 3002–3014, Nov. 2021.

[33] J. Liu, Y. Sun, W. Gan, X. Xu, B. Wohlberg, and U. S. Kamilov, "SGD-Net: Efficient model-based deep learning with theoretical guarantees," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 598–610, 2021.

[34] S. Kazemi, B. Yonel, and B. Yazici, "Unrolled Wirtinger flow with deep decoding priors for phaseless imaging," *IEEE Trans. Comput. Imag.*, vol. 8, pp. 609–625, 2022.

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

[36] J. Liu, X. Xu, W. Gan, S. Shoushtari, and U. S. Kamilov, "Online deep equilibrium learning for regularization by denoising," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 25363–25376.

[37] W. Gan et al., "Self-supervised deep equilibrium models with theoretical guarantees and applications to MRI reconstruction," *IEEE Trans. Comput. Imag.*, vol. 9, pp. 796–807, 2023.

[38] S. Welker, T. Peer, H. N. Chapman, and T. Gerkmann, "Deep iterative phase retrieval for ptychography," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, 2022, pp. 1591–1595.

[39] A. Saha, S. S. Khan, S. Sehrawat, S. S. Prabhu, S. Bhattacharya, and K. Mitra, "LWGNet-learned Wirtinger gradients for fourier ptychographic phase retrieval," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 522–537.

[40] B. Mildenhall, P.P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *Proc. Eur. Conf. Comput. Vis*, 2020, pp. 405–421.

[41] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc Med Image Comput. Comput. Assist. Interv.*, 2015, pp. 234–241.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Jan. 2017, *arXiv:1412.6980*.