Unsupervised Deep Learning for Phase Retrieval via Teacher-Student Distillation

Yuhui Quan1,2,*, Zhile Chen1,2,*, Tongyao Pang3, Hui Ji3

1School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China
2Pazhou Lab, Guangzhou 510320, China
3Department of Mathematics, National University of Singapore, 119076, Singapore
csyhquan@scut.edu.cn, cszhilechen@mail.scut.edu.cn, matpt@nus.edu.sg, matjh@nus.edu.sg

Abstract
Phase retrieval (PR) is a challenging nonlinear inverse problem in scientific imaging that involves reconstructing the phase of a signal from its intensity measurements. Recently, there has been an increasing interest in deep learning-based PR. Motivated by the challenge of collecting ground-truth (GT) images in many domains, this paper proposes a fully-unsupervised learning approach for PR, which trains an end-to-end deep model via a GT-free teacher-student online distillation framework. Specifically, a teacher model is trained using a self-expressive loss with noise resistance, while a student model is trained with a consistency loss on augmented data to exploit the teacher’s dark knowledge. Additionally, we develop an enhanced unfolding network for both the teacher and student models. Extensive experiments show that our proposed approach outperforms existing unsupervised PR methods with higher computational efficiency and performs competitively against supervised methods.

Introduction
Phase retrieval (PR) refers to reconstructing the phase of a signal from its intensity measurements, which finds a wide range of applications in scientific imaging. Formally, PR requires solving a nonlinear ill-posed problem as follows:

\[ y = |Ax_{gt}| + n, \]

where \( x_{gt} \in \mathbb{C}^N \) denotes the signal (image) to reconstruct, \( y \in \mathbb{R}^M \) the intensity measurements, \( n \in \mathbb{R}^M \) the measurement noise, \( \cdot \) the element-wise modulus operator, and \( A \in \mathbb{C}^{M \times N} \) some complex-valued linear transform, e.g., discrete Fourier transform (DFT).

In recent years, deep learning (DL) has emerged as one promising tool for PR. Most existing studies leverage supervised DL and train an end-to-end neural network (NN) over a paired set of ground-truth (GT) images and their intensity measurements. Recent works (Cha et al. 2021; Zhang et al. 2021b) can train an NN using an unpaired set. Plug-and-Play (PnP) approaches use a pre-trained denoising NN (Metzler et al. 2018; Wu et al. 2019; Shi, Lian, and Chang 2020; Wei et al. 2020; Chen et al. 2022b) or a pre-trained generative NN (Hand, Leong, and Voroninski 2018; Shamshad and Ahmed 2020; Hyder et al. 2019; Liu, Ghosh, and Scarlett 2021; Liu et al. 2021) to regularize the prediction.

These DL-based PR methods require the acquisition of a large number of GT images or images with high signal-to-noise ratio (SNR). In many domains, capturing such latent images is expensive or even infeasible. Although PnP methods can use GT images from other domains, their generality is often limited by domain shifts, e.g., statistical priors learned from digital photographs of natural scenes are not suitable for the images of biology or material science. Another approach to avoid collecting GT images is using an untrained convolutional NN (CNN) for regularizing the prediction, which is based on the deep image prior (DIP) (Ulyanov, Vedaldi, and Lempitsky 2018). These DIP-based unsupervised DL methods (Jagatap and Hegde 2019; Bostan et al. 2020; Chen et al. 2022a) have high computational cost, as different NN models need to be learned for different test samples. In addition, their performance is not as good as the supervised DL methods in the existing literature, probably due to their lack of knowledge from external data.

Motivated by the challenge of collecting GT images and the limitations of existing dataset-free unsupervised methods, this paper developed a fully-unsupervised end-to-end DL approach for PR with the following three features:

• No prerequisite on GT images or pre-trained NN models.
• Training a universal end-to-end NN from external data.
• Providing competitive performance against existing supervised DL-based methods.

Main Idea and Contributions
Unsupervised DL of PR can be interpreted as a weakly semi-supervised learning problem. Consider the image acquisition process in (1). A GT image \( x_{gt} \) is composed by a part \( x_{gt}^l \), which is completely captured by the intensity measurements \( |Ax_{gt}| \) and the other part \( x_{gt}^u \), which is completely lost during image acquisition. These two parts are sufficient for reconstructing \( x_{gt} \). Since \( y \) is a noisy version of \( |Ax_{gt}| \), it is used as a noisy label of \( x_{gt}^l \) for weakly supervised learning. However, there is no label regarding \( x_{gt}^u \). Then, the training data for unsupervised PR can be viewed as having partial weak (noisy) labels encoded by \( y \), where weakly semi-supervised learning applies.
The interpretation outlined above served as inspiration for our development of a teacher-student DL approach for unsupervised PR. Teacher-student learning is a promising semi-supervised learning technique, in which teacher models are trained on labeled data to provide initial predictions, and student models are trained to mimic the predictions of the teachers on augmented data for the purposes of regularization and improvement; see e.g. (Tarvainen and Valpola 2017; Tang et al. 2021). It is also known as online self-supervised knowledge distillation in existing literature (Anil et al. 2018; Wang and Yoon 2021), i.e., teacher and student models are jointly end-to-end trained for knowledge refinement and mutual improvement.

In the proposed approach, a teacher model employs a self-expressive loss with noise resistance to learn the prediction about $x_{gt}^l$. Such a self-expressive loss is extended from R2R loss functions for self-supervised Gaussian denoising (Pang et al. 2021) and self-supervised compressive sensing (Quan et al. 2022). Together with the image prior from the inductive bias of a deep NN (Tayal et al. 2020; Manekar et al. 2020b; Dittmer et al. 2020), the proposed loss can train the teacher model to have a reasonable prediction accuracy.

For improvement, a student model is trained together with the teacher model using a consistency loss to encourage the predictions of the student model match that of the teacher model, i.e. the so-called knowledge distillation. The consistency loss is measured on a set of paired samples formed by the image estimates from the teacher model, with data augmentation via noise injection, image transformation, and exploitation of intermediate estimates from an unfolding NN.

The motivation of using consistency learning for the student model is two-fold. Firstly, consistency learning is an effective semi-supervised learning technique (Hendrycks et al. 2019; Englesson and Azizpour 2021) for improving the noise robustness of the model. Secondly, knowledge distillation can exploit the dark knowledge from the teacher model via implicit ensemble (Allen-Zhu and Li 2020). It is observed that the samples generated through data augmentation for student training contain multiple diverse estimates of each GT image patch. This corresponds to one type of dark knowledge. Due to the weak supervision provided by its training data, the teacher model may suffer from overfitting, resulting in large prediction variance. Then, the ensemble of the multiple estimates of each GT image can significantly reduce this variance. By training the student model to predict all these estimates, it implicitly learns to perform effective ensemble. Furthermore, through consistency learning, the teacher and student models can integrate their different predictions to reduce solution ambiguity.

The performance of a DL method is greatly influenced by the NN architecture. Based on proximal gradient, we implement a deep unfolding NN for both the teacher and student models. For further improvement, we introduce two modules. One is a condition-aware module for training a universal model that adapts to imaging conditions (e.g., noise level and compression ratio). The other is a long short-term memory (LSTM) module to form a highway across the NN for more efficient feature delivery. This is the first time that memory is integrated into an unfolding NN for PR. These two modules result in a powerful NN for PR.

The performance of our proposed approach is extensively evaluated under various settings. The results indicate that our proposed approach outperforms existing GT-free methods by a large margin in terms of reconstruction accuracy and achieves competitive performance compared to the latest GT-based methods. Additionally, our approach has advantages in terms of computational complexity when compared to DIP-based unsupervised methods. The main technical contributions of this paper are as follows:

- The first work on end-to-end fully-unsupervised (GT-free) deep learning for PR with noisy measurements.
- A self-supervised loss with noise resistance for teacher model and a distillation scheme for student model.
- A deep unfolding NN enhanced for PR.

Related Works

Traditional PR methods impose some prior on images to regularize the process. One often-used prior is the sparsity prior of an image in some transform (dictionary), which results in some $\ell_1$-regularized models; see e.g. (Tillmann, Eldar, and Mairal 2016; Qiu and Palomar 2017; Chang et al. 2018; Shi et al. 2018a; Shi, Lian, and Fan 2019). Patch recurrence is another popular image prior which is often implemented by including a non-local denoiser in an iterative PR method; see e.g. (Metzler, Maleki, and Baraniuk 2016; Shi et al. 2018b).

In recent years, there is an increasing interest in DL for PR. The supervised methods (Rivenson et al. 2018; Igü, Oktem, and Koç 2019; Naimipour, Khobahi, and Soltanalian 2020; Hyder, Cai, and Asif 2020; Zhang et al. 2021a; Shi and Lian 2022) train an end-to-end NN over a paired dataset. Most of them adopt an unfolding NN which replaces the regularization-related steps in an iterative scheme by learnable modules. Based on some pre-trained deep denoising NN models, Wei et al. (2020) proposed an end-to-end NN with reinforcement learning blocks to predict the hyper-parameters involved in an unfolded scheme. The CycleGAN (Zhu et al. 2017) inspired DL methods (Cha et al. 2021; Zhang et al. 2021b) weaken the prerequisite on training data, from paired samples to the unpaired ones. Cha et al. (2021) developed a PhaseCut-based loss for improving generator training. Zhang et al. (2021b) introduced the imaging physics and a Fourier loss to improve cycle learning.

Instead of end-to-end training, PhP methods utilize pre-trained models from other image domains for regularization. Many PhP methods incorporate deep denoisers pre-trained on noisy/clean image pairs into an unfolding NN. Metzler et al. (2018) unfolded the RED (Romano, Elad, and Milanfar 2017) with the FASTA algorithm (Goldstein, Studer, and Baraniuk 2014) and plugged the pre-trained DnCNN models (Zhang et al. 2017). Chen et al. (2022) plugged pre-trained complex-valued NNs into RED. Shi, Lian, and Chang (2020) unfolded a sparse model embedded with a well-designed PhP denoiser. The success of these methods depends on how correlated the images for pre-training are to
target images. Another PNP approach leverages deep generative model for regularization; see e.g. (Hand, Leong, and Voroninski 2018; Shamshad and Ahmed 2020; Hyder et al. 2019; Liu, Ghosh, and Scarlett 2021; Liu et al. 2021). The latent image is represented by a pre-trained generative model with a specific input code. Then, the NN is trained by optimizing the code to fit the measurements. As generative models are domain-specific, these methods usually do not generalize well to unseen domains.

Dataset-free DL for PR has recently made significant progress; see e.g. (Jagatap and Hegde 2019; Bostan et al. 2020; Sun and Bouman 2021; Wang et al. 2020; Chen et al. 2022a; Wang, Li, and Ji 2022). These methods adjust the weights of an untrained NN to match the measurements of the test image, where noise sensitivity and solution ambiguity are partially addressed by the DIP induced by a CNN.

This paper is one of the few studies on unsupervised end-to-end for PR on external dataset without GT images. It effectively avoids the issues present in existing DIP-based dataset-free DL methods. Prior studies, e.g., (Tayal et al. 2020; Manekar et al. 2020b,a), have shown the inherent difficulty of learning on images with strong symmetry. Their recipes are based on some regularized reconstructive losses which work well only for noiseless intensity measurements. In contrast, our approach considers measurement noise and improves upon the issue of solution ambiguity through the use of teacher-student distillation.

Methodology

Given a set of measurement samples, but without their GT images, the goal is to train an end-to-end NN for reconstruction. Each measurement sample \( y \) relates to its GT \( x_{gt} \) by \( y = |Ax_{gt}| + n \), with measurement noise \( n \). Our main idea discussed in previous sections is implemented as follows.

Teacher-Student Distillation Framework

The teacher-student self-supervised knowledge distillation framework of the proposed approach is outlined in Figure 1. There are a teacher model \( f_T \) and a student model \( f_S \), which are simultaneously end-to-end trained. The teacher model will prepare multiple estimates of the latent images with reasonable accuracy by some scheme, and these estimates will be passed to the student for knowledge distillation. Afterward, the student model will be used for testing.

Noise-resistant learning of teacher model

The teacher model is trained through partial weak supervision provided by the measurements themselves. There are two issues to address: the existence of measurement noise and the solution ambiguity caused by the missing phase in training samples.

To address the issue caused by measurement noise, the following self-supervised loss is introduced:

\[
\mathcal{L}_T = \mathbb{E}_y \| |Af_T(y + \eta)| - (y - \eta) \|_2^2,
\]

where \( \eta \) is drawn from \( \mathcal{P}_n \), the distribution of \( n \). The loss is defined on a pair of re-corrupted measurements \( (y + \eta, y - \eta) \), and the rationale comes from the following proposition.

Proposition 1. Suppose \( \eta \) and \( n \) are independent and identically distributed (i.i.d.). Then

\[
\mathbb{E}_y \mathcal{L}_T = \mathbb{E}_{x_{gt}, n, \eta} \| Af_T(y + \eta) - Ax_{gt} \|_2^2 + \text{const..}
\]

Proposition 1 states that \( \mathcal{L}_T \) is immune to measurement noise, as it provides an unbiased estimate of the loss defined on the noise-free measurements \( |Ax_{gt}| \). In other words, the loss \( \mathcal{L}_T \) can effectively remove the negative impact caused by \( n \). However, it cannot remove other solution ambiguity, as it only considers the fitting error on intensity measurements. However, as a CNN architecture is likely to have a good inductive bias for natural images (Tayal et al. 2020; Manekar et al. 2020b,a; Dittmer et al. 2020), the teacher model can alleviate solution ambiguity with the loss \( \mathcal{L}_T \).

Consistent learning of student model

During learning, the teacher model prepares a set of image estimates \( \{x'\}_x \), as follows. (i) Noise injection: For each measurement \( y \), the teacher model takes \( y' := y + z \) as input with randomly added \( z \sim \mathcal{P}_n \), and outputs multiple image estimates \( x' = f_T(y') \). (ii) Intermediate reusing: As the teacher model is an unrolling NN with multiple stages where each stage outputs an intermediate estimate, we take the estimates from the last \( N \) stages as \( x' \), which provides various corrupted versions of latent images. (iii) Image data augmentation: we apply random rotation and random cropping on those image estimates to enlarge the set of \( x' \). Flipping is not used, as it will generate images with the same measurements and training the student model to predict different images from the same measurements will lead to contradiction.

The three schemes above enable the teacher model to have multiple estimates of the target image from different perspectives. Afterward, many pairs of training samples \( \{(y'', x')\}_x = \{Ax_{gt} + n'\}_x \sim \mathcal{P}_n \), are formed and used to train student model for consistency regularization and knowledge distillation via the following loss:

\[
\mathcal{L}_S = \mathbb{E}_{y''} \| f_s(y'') - x' \|_2^2.
\]

Total training loss

We impose the losses \( \mathcal{L}_T, \mathcal{L}_S \) on the output of every stage of the unfolding NNs, which are denoted by \( \mathcal{L}_T^k, \mathcal{L}_S^k \) respectively. Then, the teacher and student models are jointly trained by

\[
\mathcal{L} := \lambda_T \sum_{k=1}^{K} \gamma_k \mathcal{L}_T^k + \lambda_S \sum_{k=1}^{K} \gamma_k \mathcal{L}_S^k,
\]

where \( \lambda_T, \lambda_S \in \mathbb{R}^+ \) and \( \gamma_k = 1/(K - k + 1) \).

Network Architecture

Similar to standard CNNs, an unfolding NN based on proximal gradient descend also has an inductive bias (Dittmer et al. 2020).
et al. 2020) to facilitate unsupervised DL. Therefore, for both the teacher and student models, we construct an NN with K stages via unfolding the proximal gradient descend solver (Combettes and Pesquet 2011) for a regularized variational problem: $\min \|y - |Ax|\|_2^2 + \phi(x)$. Starting from an initial point $x_0$, the proximal gradient descend iterates:

$$z_k = x_{k-1} - q_k \nabla D(x_{k-1}; y, A),$$

$$x_k = \text{Prox}_{\phi}^{q_k}(z_k) := \arg\min_{x} \{\phi(x) + \frac{q_k}{2} \|x - z_k\|_2^2\},$$

where $q_k \in \mathbb{R}^+$ is a step size, $D(x; y, A) = \|y - |Ax|\|_2^2$, and $\text{Prox}_{\phi}^{q_k}(z)$ denotes the proximal operator. We replace $\text{Prox}_{\phi}^{q_k}$ by a so-called proximal module (PM) without weight sharing across stages, which is a U-Net with two enhancements. See Figure 2 for the resulting NN architecture.

**Imaging condition awareness** Imaging conditions such as noise level and sampling ratio can vary for different samples. Many existing methods, e.g. (Metzler, Maleki, and Baraniuk 2016; Metzler et al. 2018; Wei et al. 2020; Yang et al. 2022), include them as known hyper-parameters or an additional input. Instead, we introduce a condition-aware block (CAB) to utilize imaging conditions for better prediction, which also allows training a single model that generalizes well on the samples with varying imaging conditions. Let $\theta = [|\beta|, \rho]$ store the noise level $|\beta|$ (e.g., standard deviation for Gaussian noise and strength for Poisson noise) and the sampling ratio $\rho$. The CAB is a stack of fully-connected layers, which maps $\theta$ to the step sizes $\{q_k\}_{k=1}^{K}$ used in (6) as well as to a set of feature values $\{p_k\}_{k=1}^{K}$ incorporated into the PMs at different stages. Concretely, $p_k$ is repeated to form a map of the size of $x$ and used as the additional input of the $k$th PM.

**Cross-stage feature delivery** In a PM, an input image is mapped to features and then transformed back into an image for output. Then, an unrolling NN constructed via (6) alternates between the image and feature domains. Since the PMs at different stages play a similar role (i.e. proximal operators), their extracted features should be highly correlated and the features from the previous PM could benefit the process of next one. However, the aforementioned features-image-features pipeline is not efficient which may form a bottleneck for feature delivery through the whole NN, particularly when the image size is much less than the feature size.

To address the bottleneck issue of feature delivery, similar to the work for compressed sensing (Song, Chen, and Zhang 2021), we introduce convolutional long short-term memory (ConvLSTM) cells (Shi et al. 2015) on top of the CNNs, which creates a path that allows interactions and feature delivery across different stages. Then, the pipeline of our unrolling NN reads as follows: for $k = 1, \cdots, K$,

$$z_k = x_{k-1} - q_k \nabla D(x_{k-1}; y, A),$$

$$x_k = \text{PM}_k(p_k, h_{k-1}, z_k),$$

$[h_k, c_k] = \text{ConvLSTM}(t_k, h_{k-1}, c_{k-1}),$ if $k \leq K - 1,$

where $q_k, p_k$ are the output of CAB, $t_k$ is the intermediate features drawn in the $k$th PM (i.e. PM$k$), and $h_k, c_k$ are the hidden and cell states respectively in the $k$th ConvLSTM cell, with $h_0$ and $c_0$ set to zeros. To fully exploit feature delivery, we draw intermediate features from three parts of PM$k$ to form $t_k$, as shown in Figure 2. We insert a ConvLSTM cell between every two adjacent stages so as to reduce the cell number from $K$ to $K - 1$, which differs from (Song, Chen, and Zhang 2021). In addition, a ConvLSTM cell in our NN consumes features from multiple layers, rather than a single layer like (Song, Chen, and Zhang 2021).

**Experiments**

Performance evaluation is conducted on three types of measurements: coded diffraction patterns (CDPs), holographic patterns, and pychographic patterns. Through the experiments, we set $K = 5$ for both NNs and $\lambda_1 = \lambda_2 = 1, N = 2$ for training. The teacher and student models are jointly trained using the Adam optimizer with 200 epochs and batch size of 8. The learning rate is initialized to $5 \times 10^{-4}$ when the measurement number is two times larger than the pixel number, and $1 \times 10^{-3}$ otherwise. It is then decayed every 100 epochs with the factor of 0.5. To simulate practical scenarios, for each GT image, only one intensity measurement sample is generated to form the data for unsupervised learning. The trained student model is used for inference.

**Evaluation on Coded Diffraction Patterns**

CDPs in coded diffraction imaging are generated with $A = [(FD_1)^T, \cdots, (FD_J)^T]^T$, where $F$ is a DFT matrix, and $D_1, \cdots, D_J$ are defined as $D_jx \rightarrow d_j \odot x, j = 1, \cdots, J.$ Here $\odot$ denotes the Hadamard product, and $d_j \in \mathbb{C}^N$ is an illumination mask set to uniform masks or bipolar masks: the former for non-compressive CDPs and the both for compressive CDPs. A uniform mask is generated through drawing its elements uniformly from the cell circle in the complex plane. A bipolar mask is generated through drawing its elements from $\{1, -1\}$ with the Bernoulli distribution $B(1/2)$. The mask number is set to $J = 1, 2, 4$ respectively.

**PR from non-compressive uniform CDPs**

The training data setting for PR varies in existing works. Following the representative work (Wei et al. 2020), our training set consists of the 400 images of the Berkeley segmentation dataset (BSD) (Martin et al. 2001) and the 5600 images selected randomly from the PASCAL VOC dataset (Everingham et al. 2015). Each of these 6000 images is resized to $128 \times 128$ and used to generate the CDPs via (1), with Poisson noise simulated by the scheme of (Metzler et al. 2018): $y^* = |Ax|^2 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \gamma^2 \text{Diag}(|Ax|^2))$ and
Table 1: Quantitative results on uniform CDPs in terms of PSNR (dB). The best and second best results at each row are boldfaced and underlined respectively. Left part of methods: GT-dependent; Right part of methods: GT-free.

<table>
<thead>
<tr>
<th>J</th>
<th>γ</th>
<th>prDeep</th>
<th>PPR</th>
<th>DPSR</th>
<th>TFPnP</th>
<th>prCom</th>
<th>Dolpin</th>
<th>B-GAMP</th>
<th>conPR</th>
<th>DDec</th>
<th>DMMSE</th>
<th>E2E</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>35.29</td>
<td>33.29</td>
<td>33.33</td>
<td>36.05</td>
<td>35.59</td>
<td>26.29</td>
<td>35.10</td>
<td>32.81</td>
<td>33.33</td>
<td>34.04</td>
<td>30.31</td>
<td>35.94</td>
</tr>
<tr>
<td>27</td>
<td>35.34</td>
<td>32.83</td>
<td>30.15</td>
<td>29.75</td>
<td>28.19</td>
<td>27.22</td>
<td>26.93</td>
<td>28.29</td>
<td>28.93</td>
<td>29.38</td>
<td>34.10</td>
<td>32.16</td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>22.08</td>
<td>24.04</td>
<td>23.92</td>
<td>24.12</td>
<td>23.52</td>
<td>17.47</td>
<td>22.96</td>
<td>20.44</td>
<td>22.55</td>
<td>23.83</td>
<td>17.24</td>
<td>25.00</td>
<td></td>
</tr>
</tbody>
</table>

| prDeep12 | 9   | 37.61  | 35.92 | 35.90 | 38.53 | 38.06 | 30.89  | 37.75  | 35.09 | 33.21 | 37.01 | 34.27 | 38.58 |
| 27 | 31.26  | 30.63  | 30.66 | 32.07 | 32.15 | 25.69 | 30.96  | 29.48  | 28.71 | 30.06 | 24.66 | 32.16 |

| BSD68 | 9   | 34.83  | 33.46 | 33.44 | 35.46 | 34.37 | 27.43  | 34.62  | 32.98 | 32.24 | 33.97 | 27.69 | 34.23 |
| 27 | 25.92  | 28.59  | 28.75 | 29.88 | 28.75 | 25.54 | 29.05  | 26.93  | 28.11 | 29.67 | 24.41 | 29.85 |

| Time(sec.) | 9.05 | 1.72  | 5.13  | 0.02  | 7.56  | 10.09 | 16.96  | 3.61  | 22.18 | 267  | 0.02  | 0.02  |

Table 1: Quantitative results on uniform CDPs in terms of PSNR (dB). The best and second best results at each row are boldfaced and underlined respectively. Left part of methods: GT-dependent; Right part of methods: GT-free.

a larger $\gamma$ indicates a lower SNR. The value of $\gamma$ is uniformly sampled from $\{9, 27\}$. For testing, the images of the prDeep12 dataset (Metzler et al. 2018) and BSD68 (Martin et al. 2001) are used for generating measurements, corrupted by Poisson noise of $\gamma = 9, 27, 81$ respectively.

Totally eleven methods are selected for performance comparison, including DOLPHIn (Tillmann, Eldar, and Mairal 2016), B-GAMP (Metzler, Maleki, and Baraniuk 2016), ConPR (Shi et al. 2018b), prDeep (Metzler et al. 2018), PPR (Shi, Lian, and Fan 2019), DDec (Jagatap and Hegde 2019), DPSR (Shi, Lian, and Chang 2020), E2E (Manekar et al. 2020a) TFPnP (Wei et al. 2020), prCom (Chen et al. 2022b) and DMMSE (Chen et al. 2022a). Their results are quoted from (Chen et al. 2022b,a) whenever possible and otherwise obtained with their published codes. Specifically, DOLPHIn, B-GAMP and ConPR are learning-free methods, DDec and DMMSE are DIP-based dataset-free unsupervised methods, and E2E is a dataset-based unsupervised method. All of them are GT-free. In comparison, prDeep, PPR, DPSR, prCom and TFPnP are GT-dependent: the former four are PnP methods, and the last one is an end-to-end DL-based method. For fair comparison and for PSNR improvement, we replace the U-Net used in E2E by ours.

See Table 1 for the quantitative results of all compared methods measured by Peak-Signal-to-Noise Ratio (PSNR). Among all GT-free methods, ours is the best performer. By leveraging teacher-student learning, our approach noticeably outperformed the very recent DIP-based unsupervised method DMMSE and the very recent PnP method prCom. It also performs much better than the dataset-based unsupervised method E2E. Surprisingly, it even performed competitively against the representative supervised method TFPnP, with better results in more than one half settings. See Figure 3 for visual comparison on some reconstructed images. The visual quality of our results is competitive against that of the supervised methods. All above results have demonstrated the effectiveness of our approach.

**Computational efficiency** Table 1 also lists the inference time of different methods in reconstructing a $128 \times 128$ image with a uniform mask, run on an RTX Titan GPU. The TFPnP, E2E, DDec, DMMSE and our model are all implemented with PyTorch. It can be seen that our model, E2E and TFPnP have nearly the same inference time which is much less than the DIP-based methods DDec and DMMSE. This showed the computational efficiency advantage of our dataset-based unsupervised approach over the dataset-free ones. Other methods are not implemented with PyTorch and require more stages/iterations in their processing, and their running time is much higher than ours.

**PR from compressive uniform or bipolar CDPs** Compressive PR considers an additional compression process during measuring, which can be expressed as $A = CFD$, where $C$ is a $M \times N$ matrix produced by randomly sampling $M$ rows of an $N \times N$ identity matrix, and $D$ is a diagonal matrix associated with a single illumination mask. The sampling ratio $\rho$ is defined as $\rho = M/N$. Two types of noise are considered respectively: the Poisson noise simulated by the scheme of (Metzler et al. 2018) with strength $\gamma$ and the additive white Gaussian noise (AWGN) with strength measured by the SNR of input measurements. Three sampling ratios $\rho = 0.3, 0.4, 0.5$ are used respectively.

Following Yang et al. (2022), we randomly crop 6000 patches of size $128 \times 128$ from 400 images of BSD to generate the measurements for training. The Poisson noise
of training data has its strength uniformly sampled from [0,50], while the AWGN has its SNR uniformly sampled from [10,50]. For test, the images from prDeep12 are used as GTs. The strength of Poisson noise is set to 10 and 30 respectively, and the SNR of AWGN is set to 10 and 20 respectively. A single model is trained for dealing varying $\rho$.

Three baseline methods, B-GAMP, prDeep and PPR, are used for comparison. In addition, the DPUNet (Yang et al. 2022), a supervised end-to-end NN, is introduced for the comparison on uniform masks, with quoted results. See Table 2 for the quantitative results. Our approach is the top performer through all settings except for two cases where it performed slightly worse than two GT-based methods, DPUNet and PPR, respectively. Such results again demonstrated the effectiveness of our approach.

**Evaluation on Other Patterns**

**PR from holographic patterns** In holography, the measurements are generated with $A : x \rightarrow [(Fx)\mathbf{T}, (F(x + \mathbb{G}^{s_{1},s_{2}}x))\mathbf{T}, (F(x - i\mathbb{G}^{s_{1},s_{2}}x))\mathbf{T}]^\mathbf{T}$, where $(\mathbb{G}^{s_{1},s_{2}}x)(t_{1} + t_{2}n_{1}) = \exp(\frac{2\pi is_{1}t_{1}}{n_{1}} + \frac{2\pi is_{2}t_{2}}{n_{2}})x(t_{1} + t_{2}n_{1}), 0 \leq t_{j} \leq n_{j} - 1$ for $j = 1, 2$, and $i = \sqrt{-1}$. Both the $s_{1}$ and $s_{2}$ are set to 0.5 according to (Chang et al. 2018). Poisson noise defined by $|y| \sim \alpha \cdot \text{Poisson}(|Ax|/\alpha)$ is added where a larger $\alpha$ indicates a lower SNR. We randomly select 6000 (100) images from the public fashion product image dataset (Aggarwal 2019) to form the training (test) set. All the images are converted to gray-scale, resized and cropped centrally to $128 \times 128$ for generating the measurements. The Poisson noise strength $\alpha$ is uniformly sampled from $\{3, 9\}$ for training data, and set to 3 and 9 respectively on test data.

**PR from ptychographic patterns** Ptychography is one application of PR. Following (Chang et al. 2018), we define $A : x \rightarrow [(F(\omega \otimes R_{1}x))\mathbf{T}, (F(\omega \otimes R_{2}x))\mathbf{T}, \cdots, (F(\omega \otimes R_{L}x))\mathbf{T}]^\mathbf{T}$, where $R_{l}$ is a binary matrix that selects a window of $x$ and $L = 9$, and $\omega$ denotes the coded pattern generated by a $64 \times 64$ zone plate len. We se-
lect 82 (20) cell images from public microscope cell image dataset (Payavuła 2018) to form the training (test) set. The training images are cropped to 6012 patches of size 128×128 for measurement generation, and the test images are cropped to 128×128. The previous noise setting is used.

Results and analysis The prDeep (Metzler et al. 2018), PPR (Shi, Lian, and Fan 2019), DPSR (Shi, Lian, and Chang 2020), prCom (Chen et al. 2022b), and TFPnP (Wei et al. 2020) are used for comparison. To simulate the case without GT images, we directly call the models of these methods trained on CDPs of natural images. See Table 3 for the results. Our approach is the best performer. The domain gap between natural images and cell images makes most PnP methods not work well. While TFPnP can have a noticeable performance gain after retrained on cell images (denoted by TFPnP*), its result is still worse than ours. See Figure 4 for a visual comparison. Our approach preserved more structural details, while other two methods produced over-smoothing patterns or flecked backgrounds. These results have justified the value of our unsupervised learning approach for PR.

Ablation Studies
We construct the following baselines for ablation studies. (a) Teacher: using the teacher model for test; (b) w/o $L_T$: replace $L_T$ with the one used in (Manekar et al. 2020a): $\|A_f(T(y)) - y\|_2^2$; (c) w/o Inject (w/o Inter, w/o Augment): Noise injection (Intermediate reusing, image data augmentation) is disabled respectively when the teacher model prepares the data for student learning; (d) w/o CAB: all CABs are disabled; (e) w/o LSTM: all the ConvLSTM cells are replaced by a series of convolutional layers of nearly the same parameter number; (f) supervised: training the student model using paired data with MSE loss. All baselines are retrained with the same strategy stated in CDPs for fair comparison.

See Table 4 for the quantitative comparison. Each component of the proposed approach contributes to performance improvement. Particularly, (i) benefiting from the proposed noise-resistant self-expressive loss and the inductive bias of the unfolding NN architecture, the teacher model already has a not bad performance, which is further improved noticeably with larger than 0.72dB PSNR gain; (ii) the proposed noise-resistant loss $L_T$ is critical to the success of the learning; (iii) the three strategies used in the data preparation process can improve the effectiveness of student learning and distillation; (iv) our approach performed even slightly better than its supervised counterpart, which is probably due to the difficulty of supervised learning of PR on the training data with certain symmetry (Tayal et al. 2020).

Analysis on Noise Model Mismatch
Like many existing approaches, e.g. (Metzler et al. 2018; Wei et al. 2020), one key of our scheme requires the knowledge of noise. We investigate the influence of mismatch between the noise model used in our training scheme and that of both training and test data. This is done by adding Poisson-Gaussian mixed noise simulated by the scheme of (Khademi et al. 2021) to the training and test samples. The strengths of Poisson and Gaussian noise are set to 10 and 30 respectively. In training, the Poisson noise instead of the Poisson-Gaussian mixed one is used for our unsupervised loss and noise injection, with its strength $\gamma$ sampled from 9, 27, 81. As a result, the noise characteristics are inconsistent between data and our assumption. We include five selected methods for comparison. These methods are also blind to the noise characteristics of test data, with their models trained on Poisson noise directly applied.

See Table 5 for the results of CDPs on prDeep12, where our result with matched noise statistics, denoted as "Ours*", is also included. With a 0.5dB decrease in PSNR, our approach still performs better than prDeep, DPSR, DMMSE and E2E. But the performance is below that of TFPnP, as the noise robustness of TFPnP is likely to come from the image prior learned from GTs. In comparison, same as other unsupervised methods, ours does not access any GT and its noise robustness comes from the training loss whose effectiveness relies on noise model match.

Table 4: Results of ablation studies in PSNR(dB), conducted on uniform CDPs with $J=1,4$ on prDeep12 dataset, in the presence of Poisson noise with $\gamma=9$.

<table>
<thead>
<tr>
<th>$J$</th>
<th>Teacher</th>
<th>w/o $L_T$</th>
<th>w/o Inject</th>
<th>w/o Inter</th>
<th>Original</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35.21</td>
<td>35.22</td>
<td>35.62</td>
<td>35.76</td>
<td>35.94</td>
</tr>
<tr>
<td>4</td>
<td>40.29</td>
<td>40.35</td>
<td>40.72</td>
<td>40.85</td>
<td>41.09</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$J$</th>
<th>w/o Augment</th>
<th>w/o CAB</th>
<th>w/o LSTM</th>
<th>Supervised</th>
<th>Original</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35.73</td>
<td>35.34</td>
<td>35.43</td>
<td>35.90</td>
<td>35.94</td>
</tr>
<tr>
<td>4</td>
<td>40.81</td>
<td>40.66</td>
<td>40.55</td>
<td>41.02</td>
<td>41.09</td>
</tr>
</tbody>
</table>

Table 5: Performance comparison of CDPs ($J=4$) in PSNR(dB) when trained with a mismatched noise model.

<table>
<thead>
<tr>
<th></th>
<th>prDeep</th>
<th>DPSR</th>
<th>TFPnP</th>
<th>DMMSE</th>
<th>E2E</th>
<th>Ours</th>
<th>Ours*</th>
</tr>
</thead>
<tbody>
<tr>
<td>25.73</td>
<td>27.11</td>
<td>28.42</td>
<td>25.36</td>
<td>24.44</td>
<td>28.12</td>
<td>28.61</td>
<td></td>
</tr>
</tbody>
</table>

Conclusion
This paper proposed a teacher-student distillation approach to unsupervised deep learning for PR, which bypasses the difficulty of collecting GT images. The approach involves training a teacher model with a noise-resistant loss and a student model with consistent learning on paired samples generated by the teacher model. Using an unrolling NN with specific modules designed for PR, our proposed approach has been demonstrated effective and efficient in extensive experiments under various settings. In the future, we plan to explore other frameworks and strategies for teacher-student learning and distillation to further improve the approach.

Acknowledgments
Yuhui Quan acknowledges the support from Natural Science Foundation of Guangdong Province under Grant No. 2022A1515011755. Hui Ji acknowledges the support from Singapore MOE Academic Research Fund (AcRF) under Grant No. R-146-000-315-114.
References


