Unsupervised Deep Learning for Phase Retrieval via Teacher-Student Distillation

Yuhui Quan[1,2,*], **Zhile Chen**[1,2,*], **Tongyao Pang**[3], **Hui Ji**[3]

[1]School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China
[2]Pazhou Lab, Guangzhou 510320, China
[3]Department 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:

$$\boldsymbol{y} = |\mathbf{A}\boldsymbol{x}_{\text{gt}}| + \boldsymbol{n}, \tag{1}$$

where $\boldsymbol{x}_{\text{gt}} \in \mathbb{C}^N$ denotes the signal (image) to reconstruct, $\boldsymbol{y} \in \mathbb{R}^M$ the intensity measurements, $\boldsymbol{n} \in \mathbb{R}^M$ the measurement noise, $|\cdot|$ the element-wise modulus operator, and $\mathbf{A} \in \mathbb{C}^{M\times N}$ 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 $\boldsymbol{x}_{\text{gt}}$ is composed by a part $\boldsymbol{x}_{\text{gt}}^{\text{l}}$ which is completely captured by the intensity measurements $|\mathbf{A}\boldsymbol{x}_{\text{gt}}|$ and the other part $\boldsymbol{x}_{\text{gt}}^{\text{u}}$ which is completely lost during image acquisition. These two parts are sufficient for reconstructing $\boldsymbol{x}_{\text{gt}}$. Since $\boldsymbol{y}$ is a noisy version of $|\mathbf{A}\boldsymbol{x}_{\text{gt}}|$, it is used as a noisy label of $\boldsymbol{x}_{\text{gt}}^{\text{l}}$ for weakly supervised learning. However, there is no label regarding $\boldsymbol{x}_{\text{gt}}^{\text{u}}$. Then, the training data for unsupervised PR can be viewed as having partial weak (noisy) labels encoded by $\boldsymbol{y}$, where weakly semi-supervised learning applies.

[*]Corresponding authors: Zhile Chen and Yuhui Quan.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

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 $\boldsymbol{x}_{\mathrm{gt}}^{1}$. 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,a; 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.
- A self-supervised teacher-student learning approach to unsupervised PR with state-of-the-art performance.

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 ℓ_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; Işıl, 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, PnP methods utilize pre-trained models from other image domains for regularization. Many PnP 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. (2022b) plugged pre-trained complex-valued NNs into RED. Shi, Lian, and Chang (2020) unfolded a sparse model embedded with a well-designed PnP 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 $\boldsymbol{y}$ relates to its GT $\boldsymbol{x}_{\mathrm{gt}}$ by $\boldsymbol{y} = |\mathbf{A}\boldsymbol{x}_{\mathrm{gt}}| + \boldsymbol{n}$, with measurement noise $\boldsymbol{n}$. 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.

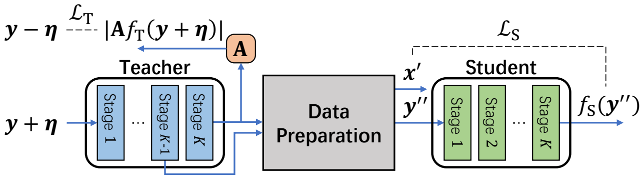
$\boldsymbol{y}-\boldsymbol{\eta}$
$\mathcal{L}_{\mathrm{T}}$
$|\mathbf{A}f_{\mathrm{T}}(\boldsymbol{y}+\boldsymbol{\eta})|$
A
Teacher
$\boldsymbol{y}+\boldsymbol{\eta}$
Stage 1
···
Stage K-1
Stage K
Data Preparation
$\boldsymbol{x}'$
$\boldsymbol{y}''$
$\mathcal{L}_{\mathrm{S}}$
Student
Stage 1
Stage 2
···
Stage K
$f_{\mathrm{S}}(\boldsymbol{y}'')$

Figure 1: Proposed framework for unsupervised DL of PR.

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:

$$\mathcal{L}_{\mathrm{T}} = \mathbb{E}_{\boldsymbol{\eta}} \| |\mathbf{A} f_{\mathrm{T}}(\boldsymbol{y}+\boldsymbol{\eta})| - (\boldsymbol{y}-\boldsymbol{\eta}) \|_2^2, \tag{2}$$

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

Proposition 1. *Suppose $\boldsymbol{\eta}|\boldsymbol{x}$ and $\boldsymbol{n}|\boldsymbol{x}$ are independent and identically distributed (i.i.d.). Then*

$$\mathbb{E}_{\boldsymbol{y}}\mathcal{L}_{\mathrm{T}} = \mathbb{E}_{\boldsymbol{x}_{\mathrm{gt}},\boldsymbol{n},\boldsymbol{\eta}} \| |\mathbf{A} f_{\mathrm{T}}(\boldsymbol{y}+\boldsymbol{\eta})| - |\mathbf{A}\boldsymbol{x}_{\mathrm{gt}}| \|_2^2 + \text{const.}. \tag{3}$$

Proposition 1 states that $\mathcal{L}_{\mathrm{T}}$ is immune to measurement noise, as it provides an unbiased estimate of the loss defined on the noise-free measurements $|\mathbf{A}\boldsymbol{x}_{\mathrm{gt}}|$. In other words, the loss $\mathcal{L}_{\mathrm{T}}$ can effectively remove the negative impact caused by $\boldsymbol{n}$. 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 $\mathcal{L}_{\mathrm{T}}$.

Consistent learning of student model During learning, the teacher model prepares a set of image estimates $\{\boldsymbol{x}'\}_{\boldsymbol{x}'}$ as follows. (i) Noise injection: For each measurement $\boldsymbol{y}$, the teacher model takes $\boldsymbol{y}' := \boldsymbol{y} + \boldsymbol{z}$ as input with randomly added $\boldsymbol{z} \sim \mathcal{P}_{\boldsymbol{n}}$, and outputs multiple image estimates $\boldsymbol{x}' = f_{\mathrm{T}}(\boldsymbol{y}')$. (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 $\boldsymbol{x}'$, 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 $\boldsymbol{x}'$. 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 $\{(\boldsymbol{y}'', \boldsymbol{x}')|\boldsymbol{y}'' = |\mathbf{A}\boldsymbol{x}'| + \boldsymbol{n}'\}_{\boldsymbol{x}',\boldsymbol{n}'\sim\mathcal{P}_{\boldsymbol{n}}}$ are formed and used to train student model for consistency regularization and knowledge distillation via the following loss:

$$\mathcal{L}_{\mathrm{S}} = \mathbb{E}_{\boldsymbol{y}''} \| f_{\mathrm{S}}(\boldsymbol{y}'') - \boldsymbol{x}' \|_2^2. \tag{4}$$

Total training loss We impose the losses $\mathcal{L}_{\mathrm{T}}, \mathcal{L}_{\mathrm{S}}$ on the output of every stage of the unfolding NNs, which are denoted by $\mathcal{L}_{\mathrm{T}}^k, \mathcal{L}_{\mathrm{S}}^k$ respectively. Then, the teacher and student models are jointly trained by

$$\mathcal{L} := \lambda_{\mathrm{T}} \sum_{k=1}^{K} \gamma_k \mathcal{L}_{\mathrm{T}}^k + \lambda_{\mathrm{S}} \sum_{k=1}^{K} \gamma_k \mathcal{L}_{\mathrm{S}}^k, \tag{5}$$

where $\lambda_{\mathrm{T}}, \lambda_{\mathrm{S}} \in \mathbb{R}^+$ 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) 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: $\min_{\boldsymbol{x}} \|\boldsymbol{y} - |\mathbf{A}\boldsymbol{x}|\|_2^2 + \phi(\boldsymbol{x})$. Starting from an initial point $\boldsymbol{x}_0$, the proximal gradient descend iterates:

$$\begin{aligned} \boldsymbol{z}_k &= \boldsymbol{x}_{k-1} - q_k \nabla \mathcal{D}(\boldsymbol{x}_{k-1}; \boldsymbol{y}, \mathbf{A}), \\ \boldsymbol{x}_k &= \mathrm{Prox}_{\phi}^{q_k}(\boldsymbol{z}_k) := \mathrm{argmin}_{\boldsymbol{x}} \{\phi(\boldsymbol{x}) + \frac{q_k}{2}\|\boldsymbol{x} - \boldsymbol{z}_k\|_2^2\}, \end{aligned} \tag{6}$$

where $q_k \in \mathbb{R}^+$ is a step size, $\mathcal{D}(\boldsymbol{x}; \boldsymbol{y}, \mathbf{A}) = \|\boldsymbol{y} - |\mathbf{A}\boldsymbol{x}|\|_2^2$, and $\mathrm{Prox}_{\phi}^{q_k}(\boldsymbol{z})$ denotes the proximal operator. We replace $\mathrm{Prox}_{\phi}^{q_k}$ 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 $\boldsymbol{\theta} = [\beta, \rho]$ store the noise level β (*e.g.*, standard deviation for Gaussian noise and strength for Poisson noise) and the sampling ratio ρ. The CAB is a stack of fully-connected layers, which maps $\boldsymbol{\theta}$ 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 $\boldsymbol{x}$ and used as the additional input of the kth PM.

Cross-stage feature delivery In a PM, an input image is mapped to features and then transformed back to 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 unfolding NN reads as follows: for $k = 1, \cdots, K$,

$$\begin{aligned} \boldsymbol{z}_k &= \boldsymbol{x}_{k-1} - q_k \nabla \mathcal{D}(\boldsymbol{x}_{k-1}; \boldsymbol{y}, \mathbf{A}), \\ \boldsymbol{x}_k &= \mathrm{PM}_k(p_k, \boldsymbol{h}_{k-1}, \boldsymbol{z}_k), \\ [\boldsymbol{h}_k, \boldsymbol{c}_k] &= \mathrm{ConvLSTM}(\boldsymbol{t}_k, \boldsymbol{h}_{k-1}, \boldsymbol{c}_{k-1}), \text{ if } k \leq K-1, \end{aligned}$$

where q_k, p_k are the output of CAB, $\boldsymbol{t}_k$ is the intermediate features drawn in the kth PM (*i.e.*, PM_k), and $\boldsymbol{h}_k, \boldsymbol{c}_k$ are the hidden and cell states respectively in the kth ConvLSTM cell, with $\boldsymbol{h}_0$ and $\boldsymbol{c}_0$ set to zeros. To fully exploit feature delivery, we draw intermediate features from three parts of PM_k to form $\boldsymbol{t}_k$, 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).

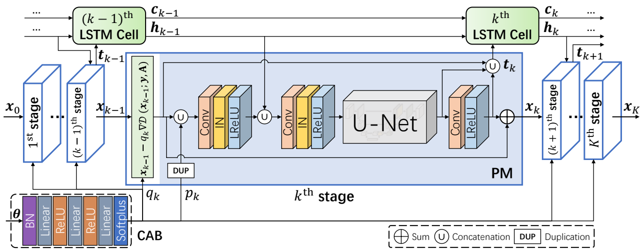

Figure 2: NN architecture for teacher and student models.

Experiments

Performance evaluation is conducted on three types of measurements: coded diffraction patterns (CDPs), holographic patterns, and ptychographic patterns. Through the experiments, we set $K = 5$ for both NNs and $\lambda_{\mathrm{T}} = \lambda_{\mathrm{S}} = 1, N = 2$ 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 5×10^{-4} when the measurement number is two times larger than the pixel number, and 1×10^{-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 $\mathbf{A} = \left[(\mathbf{F}\mathbf{D}_1)^\top, \cdots, (\mathbf{F}\mathbf{D}_J)^\top\right]^\top$, where $\mathbf{F}$ is a DFT matrix, and $\mathbf{D}_1, \cdots, \mathbf{D}_J$ are defined as $\mathbf{D}_j\boldsymbol{x} \to \boldsymbol{d}_j \odot \boldsymbol{x},\ j = 1, \cdots, J$. Here $\odot$ denotes the Hadamard product, and $\boldsymbol{d}_j \in \mathbb{C}^N$ 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 $\mathcal{B}(1/2)$. 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 128×128 and used to generate the CDPs via (1), with Poisson noise simulated by the scheme of (Metzler et al. 2018): $\boldsymbol{y}^2 = |\mathbf{A}\boldsymbol{x}|^2 + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \gamma^2 \mathrm{Diag}(|\mathbf{A}\boldsymbol{x}|^2))$ 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.

	J	γ	prDeep	PPR	DPSR	TFPnP	prCom	Dolpin	B-GAMP	conPR	DDec	DMMSE	E2E	Ours
prDeep12	1	9	35.29	33.29	33.33	**36.05**	35.59	26.29	35.10	32.81	33.33	34.04	30.31	35.94
		27	26.39	28.71	28.98	30.15	29.75	25.16	29.07	26.80	28.09	29.46	25.56	**30.17**
		81	22.08	24.04	23.92	24.42	23.52	17.47	22.96	20.44	22.55	23.53	17.24	**25.00**
	2	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**
		81	25.20	25.39	25.59	26.37	25.90	17.70	23.96	24.07	24.55	25.82	17.45	**26.40**
	4	9	39.70	37.55	37.69	40.33	40.60	31.24	40.32	36.39	37.60	40.58	37.53	**41.09**
		27	33.54	31.67	32.30	33.90	34.10	27.45	32.85	30.88	31.36	33.97	27.69	**34.23**
		81	26.90	27.02	26.73	27.23	27.60	20.22	25.43	25.87	25.19	27.12	18.58	**28.29**
BSD68	1	9	34.83	33.46	33.44	**35.46**	34.37	27.43	34.62	32.98	32.24	33.97	29.61	35.37
		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
		81	21.49	24.08	23.97	24.68	23.09	16.58	23.01	20.52	22.16	24.42	16.35	**24.99**
	2	9	37.22	35.55	35.57	37.96	37.11	32.17	37.35	35.46	32.01	36.72	34.27	**37.99**
		27	30.92	30.34	30.25	31.69	30.03	26.35	30.94	29.60	28.23	31.40	25.50	**31.69**
		81	24.70	25.35	25.38	26.28	24.94	16.54	23.98	24.10	23.73	25.54	16.59	**26.46**
	4	9	39.41	37.16	37.25	40.40	39.63	32.81	40.00	37.37	36.33	39.31	37.49	**40.52**
		27	33.14	31.53	31.87	33.63	33.19	28.52	32.82	31.05	30.67	33.12	28.64	**33.70**
		81	26.49	26.38	26.47	27.94	26.53	19.76	24.99	25.95	24.44	26.56	19.91	**28.05**
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**

a larger γ indicates a lower SNR. The value of γ 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 128×128 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 $\mathbf{A} = \mathbf{CFD}$, where $\mathbf{C}$ is a $M \times N$ matrix produced by randomly sampling M rows of an $N \times N$ identity matrix, and $\mathbf{D}$ is a diagonal matrix associated with a single illumination mask. The sampling ratio ρ 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 γ 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 128×128 from 400 images of BSD to generate the measurements for training. The Poisson noise

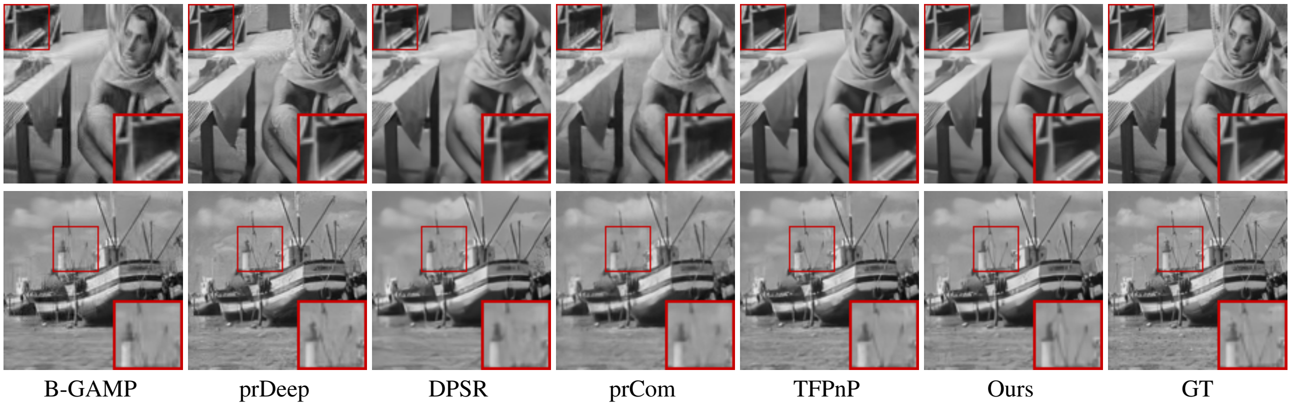

Figure 3: Reconstructed images on uniform CDPs with $\gamma = 27$. Upper: $J = 2$; Bottom: $J = 4$.

Table 2: Performance comparison of PR from compressive uniform or bipolar CDPs on prDeep12 dataset, in terms of PSNR(dB). The best result at each column is **boldfaced**.

	ρ	0.50		0.40		0.30	
Uniform	γ	10	30	10	30	10	30
	B-GAMP	32.33	27.85	31.94	27.78	30.53	27.33
	prDeep	32.43	22.29	30.90	23.80	30.55	25.10
	PPR	29.91	27.31	28.51	26.48	26.20	25.07
	DPUNet	33.18	28.63	**32.34**	28.37	30.96	27.76
	Ours	**33.19**	**28.86**	32.30	**28.52**	**31.09**	**28.12**
Bipolar	SNR	10	20	10	20	10	20
	B-GAMP	23.75	28.45	23.22	27.74	22.10	26.75
	prDeep	19.51	28.27	18.49	27.77	17.04	26.89
	PPR	**24.46**	27.17	23.69	26.16	22.72	24.33
	Ours	24.35	**28.62**	**23.71**	**27.94**	**23.13**	**26.99**

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 ρ.

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 $\mathbf{A} : \boldsymbol{x} \rightarrow [(\mathbf{F}\boldsymbol{x})^\top, (\mathbf{F}(\boldsymbol{x} + \mathscr{D}^{s_1,s_2}\boldsymbol{x}))^\top, (\mathbf{F}(\boldsymbol{x} - i\mathscr{D}^{s_1,s_2}\boldsymbol{x}))^\top]^\top$, where $(\mathscr{D}^{s_1,s_2}\boldsymbol{x})(t_1 + t_2 n_1) = \exp(\frac{2\pi i s_1 t_1}{n_1} + \frac{2\pi i s_2 t_2}{n_2})\boldsymbol{x}(t_1 + t_2 n_1), 0 \leq t_j \leq n_j - 1$ 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 $|\boldsymbol{y}| \sim \alpha \cdot \text{Poisson}(|\mathbf{A}\boldsymbol{x}|/\alpha)$ is added where a larger α 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 128×128 for generating the measurements. The Poisson noise strength α is uniformly sampled from $\{3, 9\}$ for training data, and set to 3 and 9 respectively on test data.

Table 3: PSNR(dB) results of PR from holographic patterns and ptychographic patterns respectively.

	α	prDeep	DPSR	PPR	prCom	TFPnP	TFPnP*	Ours
Holo	3	31.73	25.79	25.85	31.03	32.18	33.13	**33.81**
	9	28.70	24.88	24.96	27.93	28.52	29.58	**30.28**
Pty	3	25.37	20.93	20.93	23.99	24.46	25.12	**25.68**
	9	18.78	19.94	19.90	20.37	22.57	23.30	**24.08**

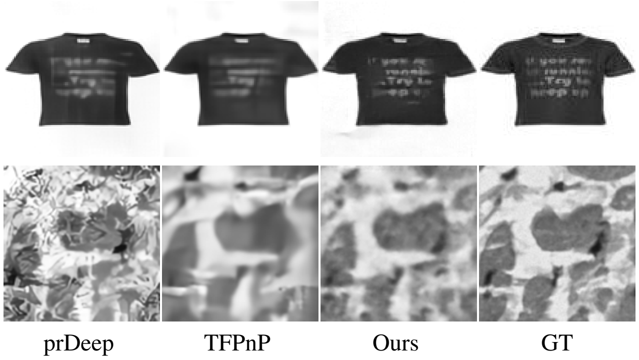

Figure 4: Reconstructed images from holographic patterns (up) and ptychographic patterns (bottom) with $\alpha = 9$.

PR from ptychographic patterns Ptychography is one application of PR. Following (Chang et al. 2018), we define $\mathbf{A} : \boldsymbol{x} \rightarrow [(\mathbf{F}(\omega \odot \mathbf{R}_1\boldsymbol{x}))^\top, (\mathbf{F}(\omega \odot \mathbf{R}_2\boldsymbol{x}))^\top, \cdots, (\mathbf{F}(\omega \odot \mathbf{R}_L\boldsymbol{x}))^\top]^\top$, where $\mathbf{R}_l$ is a binary matrix that selects a window of $\boldsymbol{x}$ and $L = 9$, and the ω denotes the coded pattern generated by a 64×64 zone plate len. We select 82 (20) cell images from public microscope cell image dataset (Payyavula 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 $\mathcal{L}_{\mathrm{T}}$: replace $\mathcal{L}_{\mathrm{T}}$ with the one used in (Manekar et al. 2020a): $\||\mathbf{A} f_{\mathrm{T}}(\boldsymbol{y})| - \boldsymbol{y}\|_2^2$; (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 $\mathcal{L}_{\mathrm{T}}$ 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).

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 $\gamma = 9$.

J	Teacher	w/o $\mathcal{L}_{\mathrm{T}}$	w/o Inject	w/o Inter	Original
1	35.21	35.22	35.62	35.76	35.94
4	40.29	40.35	40.72	40.85	41.09
J	w/o Augment	w/o CAB	w/o LSTM	Supervised	Original
1	35.73	35.34	35.43	35.90	35.94
4	40.81	40.66	40.55	41.02	41.09

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

prDeep	DPSR	TFPnP	DMMSE	E2E	Ours	Ours*
25.73	27.11	28.42	25.36	24.44	28.12	28.61

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.

Appendix

Proof of Proposition 1

Rewrite $\mathbb{E}_{\boldsymbol{y}}\mathcal{L}_{\mathrm{T}}$ by

$$\begin{aligned}\mathbb{E}_{\boldsymbol{y}}\mathcal{L}_{\mathrm{T}} = \mathbb{E}_{\boldsymbol{x}_{\mathrm{gt}},\boldsymbol{n},\boldsymbol{\eta}}\big[&\||\mathbf{A}f_{\mathrm{T}}(\boldsymbol{y}+\boldsymbol{\eta})|-|\mathbf{A}\boldsymbol{x}_{\mathrm{gt}}|\|_2^2 \\ &+2\overline{(\boldsymbol{\eta}-\boldsymbol{n})}^{\top}(|\mathbf{A}f_{\mathrm{T}}(\boldsymbol{y}+\boldsymbol{\eta})|-|\mathbf{A}\boldsymbol{x}_{\mathrm{gt}}|) \\ &+\overline{(\boldsymbol{\eta}-\boldsymbol{n})}^{\top}(\boldsymbol{\eta}-\boldsymbol{n})\big],\end{aligned} \tag{7}$$

where the last term $\mathbb{E}_{\boldsymbol{x}_{\mathrm{gt}},\boldsymbol{n},\boldsymbol{\eta}}\overline{(\boldsymbol{\eta}-\boldsymbol{n})}^{\top}(\boldsymbol{\eta}-\boldsymbol{n})$ is a constant regardless the values of the NN's parameters. Since $\boldsymbol{\eta}$ and $\boldsymbol{n}$ conditional on $\boldsymbol{x}_{\mathrm{gt}}$ are independent and follow the same distribution $\mathcal{P}_{\boldsymbol{n}}(\cdot|\boldsymbol{x}_{\mathrm{gt}})$, we can rewrite the first term of the right-hand side of Eq. (7) as follows:

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}_{\mathrm{gt}},\boldsymbol{n},\boldsymbol{\eta}}\overline{\boldsymbol{n}}^{\top}(|\mathbf{A}f_{\mathrm{T}}(\boldsymbol{y}+\boldsymbol{\eta})|-|\mathbf{A}\boldsymbol{x}_{\mathrm{gt}}|) \\ =&\,\mathbb{E}_{\boldsymbol{x}_{\mathrm{gt}}}\mathbb{E}_{\boldsymbol{n}|\boldsymbol{x}_{\mathrm{gt}}}\mathbb{E}_{\boldsymbol{\eta}|\boldsymbol{x}_{\mathrm{gt}}}\overline{\boldsymbol{n}}^{\top}(|\mathbf{A}f_{\mathrm{T}}(|\mathbf{A}\boldsymbol{x}_{\mathrm{gt}}|+\boldsymbol{n}+\boldsymbol{\eta})|-|\mathbf{A}\boldsymbol{x}_{\mathrm{gt}}|) \\ =&\int_{\boldsymbol{x}_{\mathrm{gt}}}\int_{\boldsymbol{n}|\boldsymbol{x}_{\mathrm{gt}}}\int_{\boldsymbol{\eta}|\boldsymbol{x}_{\mathrm{gt}}}P_{\boldsymbol{x}_{\mathrm{gt}}}(\boldsymbol{x}_{\mathrm{gt}})P_{\boldsymbol{n}}(\boldsymbol{\eta}|\boldsymbol{x}_{\mathrm{gt}})P_{\boldsymbol{n}}(\boldsymbol{n}|\boldsymbol{x}_{\mathrm{gt}})\overline{\boldsymbol{n}}^{\top}\cdot \\ &\qquad(|\mathbf{A}f_{\mathrm{T}}(|\mathbf{A}\boldsymbol{x}_{\mathrm{gt}}|+\boldsymbol{n}+\boldsymbol{\eta})|-|\mathbf{A}\boldsymbol{x}_{\mathrm{gt}}|) \\ =&\int_{\boldsymbol{x}_{\mathrm{gt}}}\int_{\boldsymbol{n}|\boldsymbol{x}_{\mathrm{gt}}}\int_{\boldsymbol{\eta}|\boldsymbol{x}_{\mathrm{gt}}}P_{\boldsymbol{x}_{\mathrm{gt}}}(\boldsymbol{x}_{\mathrm{gt}})P_{\boldsymbol{n}}(\boldsymbol{n}|\boldsymbol{x}_{\mathrm{gt}})P_{\boldsymbol{n}}(\boldsymbol{\eta}|\boldsymbol{x}_{\mathrm{gt}})\overline{\boldsymbol{\eta}}^{\top}\cdot \\ &\qquad(|\mathbf{A}f_{\mathrm{T}}(|\mathbf{A}\boldsymbol{x}_{\mathrm{gt}}|+\boldsymbol{n}+\boldsymbol{\eta})|-|\mathbf{A}\boldsymbol{x}_{\mathrm{gt}}|) \\ =&\,\mathbb{E}_{\boldsymbol{x}_{\mathrm{gt}},\boldsymbol{n},\boldsymbol{\eta}}\overline{\boldsymbol{\eta}}^{\top}(|\mathbf{A}f_{\mathrm{T}}(\boldsymbol{y}+\boldsymbol{\eta})|-|\mathbf{A}\boldsymbol{x}_{\mathrm{gt}}|).\end{aligned}$$

Thus, the second term of the right-hand side of Eq. (7) is zero. Also the not that the third term of the right-hand side of Eq. (7) is a constant determined by the noise's characteristics. Therefore, we have

$$\mathbb{E}_{\boldsymbol{y}}\mathcal{L}_{\mathrm{T}} = \mathbb{E}_{\boldsymbol{x}_{\mathrm{gt}},\boldsymbol{n},\boldsymbol{\eta}}\||\mathbf{A}f_{\mathrm{T}}(\boldsymbol{y}+\boldsymbol{\eta})|-|\mathbf{A}\boldsymbol{x}_{\mathrm{gt}}|\|_2^2+\mathrm{const.}.$$

The proof is done.

Details of the U-Net used in PM

The detailed structure of the U-Net used in PM is illustrated in Figure 5. Both the max pooling and transposed convolutional layers use a scaling factor of 2. All the convolutional layers are 3×3 with stride 1 and zero padding unless specified. The negative slopes in all LeakyReLU layers are 0.2.

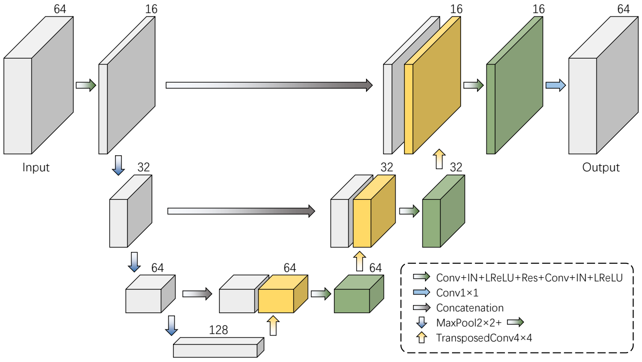

Figure 5: Detailed architecture of U-Net used in PM.

Distribution of Image Estimates from Teacher

One dark knowledge from the teacher model trained by the weak supervision is its prediction uncertainty. The data preparation process done by the teacher model indeed provides different estimates of a GT, and the student model trained to predict all such estimates indeed is learning some kind of ensemble of such estimates. Figure 6 provides an empirical study where the estimates from the teacher model via augmentations of noise injection and intermediate reusing are visualized via t-SNE, where each image estimate is plotted as a 2D point. The results show that those estimates have sufficient diversity around the GT to allow effective ensemble by the consistent learning and knowledge distillation by the student model.

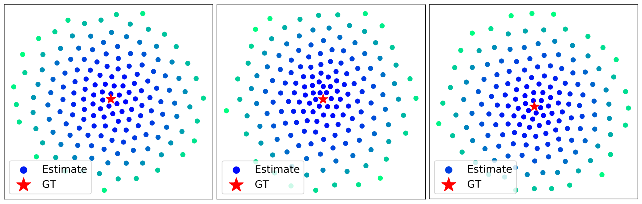

Figure 6: T-SNE visualization of image estimates from the teacher model with augmentations by noise injection and intermediate reusing. The test samples are the 1× uniform CDP measurements of three images from PASCAL VOC dataset, corrupted by Poisson noise with strength $\gamma = 9$.

Visual Comparison of Compressive CDPs

Figure 7 shows some visual results on compressive CDPs. It can be seen that E2E is not good at handling measurement noise, as the reconstructed image contains severe noise. The results of B-GAMP suffer fro over-smoothing. In contrast, the images reconstructed by our approach are much better.

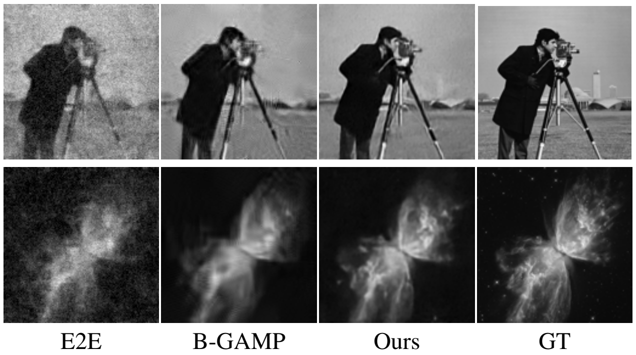

Figure 7: Images reconstructed by selected methods from compressive uniform CDP with $\rho = 0.3$ and $\gamma = 30$.

References

Aggarwal, P. 2019. Fashion Product Images Dataset. https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-dataset. Accessed: 2022-05-03.

Allen-Zhu, Z.; and Li, Y. 2020. Towards Understanding Ensemble, Knowledge Distillation and Self-Distillation in Deep Learning.

Anil, R.; Pereyra, G.; Passos, A.; Ormandi, R.; Dahl, G. E.; and Hinton, G. E. 2018. Large scale distributed neural network training through online distillation. In *International Conference on Learning Representations*.

Bostan, E.; Heckel, R.; Chen, M.; Kellman, M.; and Waller, L. 2020. Deep phase decoder: self-calibrating phase microscopy with an untrained deep neural network. *Optica*, 7(6): 559–562.

Cha, E.; Lee, C.; Jang, M.; and Ye, J. C. 2021. DeepPhaseCut: Deep Relaxation in Phase for Unsupervised Fourier Phase Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chang, H.; Lou, Y.; Duan, Y.; and Marchesini, S. 2018. Total variation based phase retrieval for poisson noise removal. *SIAM Journal on Imaging Sciences*, 11: 24–55.

Chen, M.; Peikang, L.; Quan, Y.; Pang, T.; and Ji, H. 2022a. Unsupervised Phase Retrieval Using Deep Approximate MMSE Estimation. *IEEE Transactions on Signal Processing*.

Chen, Z.; Huang, Y.; Hu, Y.; and Chen, Z. 2022b. Phase Recovery With Deep Complex-Domain Priors. *IEEE Signal Processing Letters*, 29: 887–891.

Combettes, P. L.; and Pesquet, J.-C. 2011. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, 185–212. Springer.

Dittmer, S.; Kluth, T.; Maass, P.; and Otero Baguer, D. 2020. Regularization by architecture: A deep prior approach for inverse problems. *Journal of Mathematical Imaging and Vision*, 62(3): 456–470.

Englesson, E.; and Azizpour, H. 2021. Consistency Regularization Can Improve Robustness to Label Noise. In *International Conference on Machine Learning Workshop*.

Everingham, M.; Eslami, S.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2015. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1): 98–136.

Goldstein, T.; Studer, C.; and Baraniuk, R. 2014. A Field Guide to Forward-Backward Splitting with a FASTA Implementation.

Hand, P.; Leong, O.; and Voroninski, V. 2018. Phase retrieval under a generative prior. *Advances in Neural Information Processing Systems*, 31.

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2019. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. In *International Conference on Learning Representations*.

Hyder, R.; Cai, Z.; and Asif, M. S. 2020. Solving phase retrieval with a learned reference. In *European Conference on Computer Vision*, 425–441. Springer.

Hyder, R.; Shah, V.; Hegde, C.; and Asif, M. S. 2019. Alternating phase projected gradient descent with generative priors for solving compressive phase retrieval. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 7705–7709.

Işıl, Ç.; Oktem, F. S.; and Koç, A. 2019. Deep iterative reconstruction for phase retrieval. *Applied Optics*, 58(20): 5422–5431.

Jagatap, G.; and Hegde, C. 2019. Algorithmic guarantees for inverse imaging with untrained network priors. *Advances in Neural Information Processing Systems*, 32.

Khademi, W.; Rao, S.; Minnerath, C.; Hagen, G.; and Ventura, J. 2021. Self-supervised poisson-gaussian denoising. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2131–2139.

Liu, Z.; Ghosh, S.; and Scarlett, J. 2021. Towards sample-optimal compressive phase retrieval with sparse and generative priors. *Advances in Neural Information Processing Systems*, 34.

Liu, Z.; Liu, J.; Ghosh, S.; Han, J.; and Scarlett, J. 2021. Generative Principal Component Analysis. In *International Conference on Learning Representations*.

Manekar, R.; Tayal, K.; Kumar, V.; and Sun, J. 2020a. End to end learning for phase retrieval. In *International Conference on Machine Learning Workshop*.

Manekar, R.; Zhuang, Z.; Tayal, K.; Kumar, V.; and Sun, J. 2020b. Deep learning initialized phase retrieval. In *NeurIPS 2020 Workshop on Deep Learning and Inverse Problems*.

Martin, D.; Fowlkes, C.; Tal, D.; and Malik, J. 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the Eighth IEEE International Conference on Computer Vision*.

Metzler, C. A.; Maleki, A.; and Baraniuk, R. G. 2016. BM3d-prGAMP: Compressive phase retrieval based on BM3d denoising. In *IEEE International Conference on Image Processing*, 2504–2508.

Metzler, C. A.; Schniter, P.; Veeraraghavan, A.; and Baraniuk, R. G. 2018. prDeep: robust phase retrieval with a flexible deep network. In *International Conference on Machine Learning*.

Naimipour, N.; Khobahi, S.; and Soltanalian, M. 2020. Upr: A model-driven architecture for deep phase retrieval. In *Proceedings of the 54th Asilomar Conference on Signals, Systems, and Computers*, 205–209. IEEE.

Pang, T.; Zheng, H.; Quan, Y.; and Ji, H. 2021. Recorrupted-to-Recorrupted: Unsupervised Deep Learning for Image Denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2043–2052.

Payyavula, G. 2018. Nuclei Segmentation In Microscope Cell Images. https://www.kaggle.com/datasets/gangadhar/nuclei-segmentation-in-microscope-cell-images. Accessed: 2022-05-03.

Qiu, T.; and Palomar, D. P. 2017. Undersampled sparse phase retrieval via majorization–minimization. *IEEE Transactions on Signal Processing*, 65: 5957–5969.

Quan, Y.; Qin, X.; Pang, T.; and Ji, H. 2022. Dual-Domain Self-supervised Learning and Model Adaption for Deep Compressive Imaging. In *Proceedings of the 17th European Conference on Computer Vision*, 409–426.

Rivenson, Y.; Zhang, Y.; Günaydın, H.; Teng, D.; and Ozcan, A. 2018. Phase recovery and holographic image reconstruction using deep learning in neural networks. *Light: Science and Applications*, 7(2): 17141–17141.

Romano, Y.; Elad, M.; and Milanfar, P. 2017. The little engine that could: Regularization by denoising (RED). *SIAM Journal on Imaging Sciences*, 10(4): 1804–1844.

Shamshad, F.; and Ahmed, A. 2020. Compressed sensing-based robust phase retrieval via deep generative priors. *IEEE Sensors Journal*, 21(2): 2286–2298.

Shi, B.; and Lian, Q. 2022. DualPRNet: Deep shrinkage dual frame network for deep unrolled phase retrieval. *IEEE Signal Processing Letters*.

Shi, B.; Lian, Q.; and Chang, H. 2020. Deep prior-based sparse representation model for diffraction imaging: a plug-and-play method. *Signal Processing*, 168.

Shi, B.; Lian, Q.; Chen, S.; Tian, Y.; and Xiaoyu, F. 2018a. Coded diffraction imaging via double sparse regularization model. *Digital Signal Processing*, 79: 23–33.

Shi, B.; Lian, Q.; and Fan, X. 2019. PPR: Plug-and-play regularization model for solving nonlinear imaging inverse problems. *Signal Processing*, 162: 83–96.

Shi, B.; Lian, Q.; Huang, X.; and An, N. 2018b. Constrained phase retrieval: when alternating projection meets regularization. *Journal of the Optical Society of America B: Optical Physics*, 35: 1271–1281.

Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; and Woo, W.-c. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*, 28.

Song, J.; Chen, B.; and Zhang, J. 2021. Memory-Augmented Deep Unfolding Network for Compressive Sensing. In *Proceedings of the 29th ACM International Conference on Multimedia*, 4249–4258.

Sun, H.; and Bouman, K. L. 2021. Deep probabilistic imaging: Uncertainty quantification and multi-modal solution characterization for computational imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Tang, Y.; Chen, W.; Luo, Y.; and Zhang, Y. 2021. Humble teachers teach better students for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3132–3141.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30.

Tayal, K.; Lai, C.-H.; Manekar, R.; Zhuang, Z.; Kumar, V.; and Sun, J. 2020. Unlocking inverse problems using deep learning: Breaking symmetries in phase retrieval. In *NeurIPS 2020 Workshop on Deep Learning and Inverse Problems*.

Tillmann, A. M.; Eldar, Y. C.; and Mairal, J. 2016. DOLPHIn-dictionary learning for phase retrieval. *IEEE Transactions on Signal Processing*, 64: 6485–6500.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. S. 2018. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9446–9454.

Wang, F.; Bian, Y.; Wang, H.; Lyu, M.; Pedrini, G.; Osten, W.; Barbastathis, G.; and Situ, G. 2020. Phase imaging with an untrained neural network. *Light: Science and Applications*, 9(1): 1–7.

Wang, L.; and Yoon, K.-J. 2021. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, W.; Li, J.; and Ji, H. 2022. Self-Supervised Deep Image Restoration via Adaptive Stochastic Gradient Langevin Dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1989–1998.

Wei, K.; Aviles-Rivero, A.; Liang, J.; Fu, Y.; Schnlieb, C.-B.; and Huang, H. 2020. Tuning-free plug-and-play proximal algorithm for inverse imaging problems. In *International Conference on Machine Learning*.

Wu, Z.; Sun, Y.; Liu, J.; and Kamilov, U. 2019. Online regularization by denoising with applications to phase retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

Yang, Y.; Tao, R.; Wei, K.; and Fu, Y. 2022. Dynamic proximal unrolling network for compressive imaging. *Neurocomputing*, 510: 203–217.

Zhang, F.; Liu, X.; Guo, C.; Lin, S.; Jiang, J.; and Ji, X. 2021a. Physics-based Iterative Projection Complex Neural Network for Phase Retrieval in Lensless Microscopy Imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10523–10531.

Zhang, K.; Zuo, W.; Gu, S.; and Zhang, L. 2017. Learning deep CNN denoiser prior for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3929–3938.

Zhang, Y.; Noack, M. A.; Vagovic, P.; Fezzaa, K.; Garcia-Moreno, F.; Ritschel, T.; and Villanueva-Perez, P. 2021b. PhaseGAN: A deep-learning phase-retrieval approach for unpaired datasets. *Optics Express*, 29(13): 19593–19604.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2223–2232.