A Physics-Inspired Deep Learning Framework With Polar Coordinate Attention for Ptychographic Imaging

Han Yue, Jun Cheng, *Senior Member, IEEE*, Yu-Xuan Ren, Chien-Chun Chen, Grant A. van Riessen, Philip Heng Wai Leong, *Senior Member, IEEE*, and Steve Feng Shu, *Senior Member, IEEE*

***Abstract*—Ptychographic imaging confronts inherent challenges in applying deep learning for phase retrieval from diffraction patterns. Conventional neural architectures, both convolutional neural networks and Transformer-based methods, are optimized for natural images with Euclidean spatial neighborhood-based inductive biases that exhibit geometric mismatch with the concentric coherent patterns characteristic of diffraction data in reciprocal space. In this paper, we present PPN, a physics-inspired deep learning network with Polar Coordinate Attention (PoCA) for ptychographic imaging, that aligns neural inductive biases with diffraction physics through a dual-branch architecture separating local feature extraction from non-local coherence modeling. It consists of a PoCA mechanism that replaces Euclidean spatial priors with physically consistent radial-angular correlations. PPN outperforms existing end-to-end models, with spectral and spatial analysis confirming its greater preservation of high-frequency details. Notably, PPN maintains robust performance compared to iterative methods even at low overlap ratios — well-suited for high-throughput imaging in real-world acquisition scenarios for samples with consistent structural characteristics.**

***Index Terms*—Ptychography, physics-inspired deep learning, reciprocal-space learning, transformer.**

Received 4 October 2024; revised 3 March 2025 and 18 April 2025; accepted 8 May 2025. Date of publication 6 June 2025; date of current version 2 July 2025. This work was supported by the Agency for Science, Technology and Research (A∗STAR) through MTC Programmatic Funds under Grant M23L7b0021. The associate editor coordinating the review of this article and approving it for publication was Dr. Doga Gursoy. *(Corresponding author: Steve Feng Shu.)*

Han Yue is with the Academy for Engineering & Technology, Fudan University, Shanghai 200433, China (e-mail: hyue23@m.fudan.edu.cn).

Jun Cheng is with the Institute for Infocomm Research, Agency for Science, Technology and Research (A∗STAR),, Singapore 138632 (e-mail: cheng_jun@i2r.a-star.edu.sg).

Yu-Xuan Ren is with the Institute for Translational Brain Research, Fudan University, Shanghai 200032, China (e-mail: yxren@fudan.edu.cn).

Chien-Chun Chen is with the Department of Engineering and System Science, National Tsing Hua University, Hsinchu 300044, Taiwan (e-mail: cc-chen0627@gmail.com).

Grant A. van Riessen is with the Department of Mathematical and Physical Sciences, School of Computing, Engineering and Mathematical Sciences, La Trobe University, Bundoora, VIC 3086, Australia (e-mail: g.vanriessen@latrobe.edu.au).

Philip Heng Wai Leong is with the School of Electrical and Computer Engineering, University of Sydney, Camperdown, NSW 2006, Australia (e-mail: philip.leong@sydney.edu.au).

Steve Feng Shu is with the School of Electrical and Computer Engineering, University of Sydney, Camperdown, NSW 2006, Australia (e-mail: steve.shu@sydney.edu.au).

Our code is available at: https://github.com/johncolddd/PPN

Digital Object Identifier 10.1109/TCI.2025.3572250

I. Introduction

COHERENT diffraction imaging (CDI) has enabled high-resolution, lens-less imaging across various scientific disciplines by exploiting the principles of wave propagation and interference. In CDI, detectors capture only the far-field intensity distribution of scattered coherent radiation, resulting in the loss of crucial phase information due to the well-known phase problem in crystallography. The objective of phase retrieval algorithms is to reconstruct the complete complex-valued exit wave function from these incomplete Fraunhofer diffraction patterns. Ptychography [1], an advanced CDI technique that operates in both real and reciprocal space, addresses this inverse problem by utilizing multiple overlapping diffraction measurements in reciprocal space, effectively extending the Fourier domain sampling and enabling robust phase retrieval through iterative algorithms. This approach offers extended field-of-view imaging with exceptional spatial resolution in real space. Recent breakthroughs have pushed the boundaries of resolution, achieving 0.39 Å in transmission electron microscopy [2] and even 14 pm through local-orbital ptychography [3], opening new possibilities in materials characterization [4], biological imaging [5], and semiconductor research [6].

However, the widespread application of ptychography faces significant challenges, particularly in computational efficiency. As imaging capabilities advance, data volume grows exponentially, overwhelming conventional algorithms. For instance, processing one second of data from a modern synchrotron source (10-megapixel detector, 32-bit depth, 2 kHz, 640 Gb/s) can take up to an hour [7]. Additionally, achieving high-quality retrievals requires 60-70% probe position overlap [8], further increasing computational complexity. These factors severely constrain ptychography's viability in real-time and high-throughput applications.

To address these limitations, researchers have redefined ptychography as a data-driven supervised learning task, leveraging deep learning (DL) techniques. Notable examples including PtychoNN [9], Deep-phase-imaging (DPI) [10], and PtyNet [11] have demonstrated improved retrieval efficiency. Unlike conventional approaches that require repeated iterations across the spatial and frequency domains, these methods utilize experimentally acquired diffraction patterns and complex amplitude images reconstructed by traditional algorithms as training data.

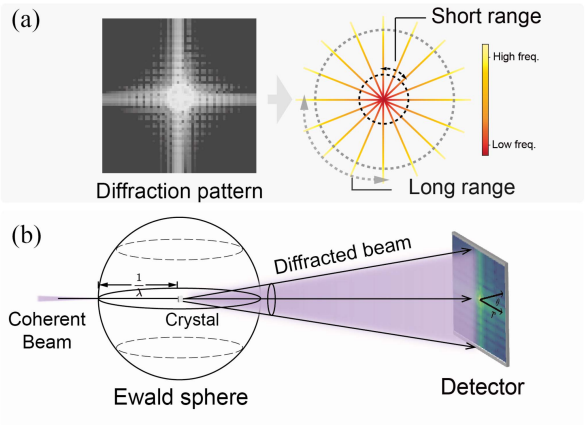

Fig. 1. Diffraction pattern characteristics and diffraction physics. (a): Diffraction patterns exhibit a radial distribution of information, with varying requirements for capturing low and high-frequency information. (b): 2D diffraction patterns can be viewed as projections of the intersection between the Ewald sphere and the crystal onto the detector plane. Sphere's radius is $1/\lambda$, the wavelength of the incident and diffracted beams. Importantly, the spatial adjacency preference observed in natural images' feature space also holds in the polar coordinate perspective of diffraction patterns.

Through convolutional neural network (CNN) based architectures, they establish a direct mapping from diffraction patterns to complex amplitude retrieval, significantly enhancing computational efficiency. Despite their success, CNN-based methods face challenges in processing multi-frequency information, capturing long-range correlations, and distinguishing high-frequency signals from noise in diffraction patterns, as illustrated in Fig. 1(a). These limitations are due to the sliding window operation of convolution kernels in spatial coordinates and the information flattening in multi-layer convolution during transmission [12], [13], [14]. Transformer models, while powerful in various computer vision tasks, have limited application in ptychography due to the mismatch between their design principles and the radial nature of diffraction patterns. For instance, Vision Transformers (ViT), are primarily optimized for natural images in real space (e.g., patch extraction operations and local feature preference priors of the Softmax function). However, these priors do not directly map to the radial nature of diffraction data in reciprocal space shown in Fig. 1(a).

Addressing the inherent limitations of existing methods requires a fundamental reconsideration of deep learning architectures for ptychographic retrieval, with a focus on optimizing neural networks for representing diffraction patterns in reciprocal space. This insight inspired the development of PPN, a novel framework that bridges frequency-domain diffraction patterns and real-space images using physics-informed deep learning techniques. Drawing inspiration from the Ewald construction in X-ray crystallography, as depicted in Fig. 1(b), treating 2D diffraction patterns as projections of the intersections between the Ewald sphere and the reciprocal lattice onto the detector plane, as shown in the left of Fig. 1. This insight revealed that diffraction patterns are inherently defined in spherical rather than planar geometry. Consequently, we developed Polar Coordinate Attention (PoCA), an attention mechanism leveraging polar coordinates that naturally aligns with diffraction physics. This polar coordinate-based attention mechanism reframes real-space structural priors in the diffraction context, mapping scattering vector magnitude to r and angular information to θ, thus capturing both radial intensity attenuation and angular coherence. Recognizing the multi-scale and multi-frequency nature of diffraction patterns, as shown in the right side of Fig. 1, PPN employs a dual-branch architecture combining a Local Dependencies Branch constructed from standard ViT blocks with a NonLocal Coherence block containing our designed PoCA. This design balances the capture of long-range and local dependencies in diffraction patterns. PPN's architecture intrinsically aligns with ptychography physics, creating a more natural correspondence between model operations and underlying physical processes.

This structural redesign based on physical insights enables our model to more effectively extract and utilize global coherent information and long-range dependencies in diffraction data, thus offering a physics-informed approach to ptychographic retrieval. Our comprehensive evaluation across simulated and real experimental datasets demonstrates PPN's multi-faceted advantages: (1) Superior retrieval quality across all metrics against leading end-to-end baselines, particularly in high-frequency preservation as evidenced through power spectral density curves and cross-sectional intensity profiles; (2) Remarkable operational viability - maintaining $<5\%$ performance degradation at 30% overlap ratio while achieving $>$ 1,000$\times$ faster inference than iterative method when tested on samples without feature distribution shifts, crucial for time-sensitive synchrotron experiments; (3) Unprecedented efficiency with 11$\times$ fewer parameters than transformer-based counterparts (6.1M vs 68.9M) and Floating Point Operations Per Second (FLOPs) comparable to lightweight CNNs, enabling deployment in real-world acquisition scenarios. These improvements could significantly enhance the applicability of ptychography in time-sensitive or radiation-sensitive imaging scenarios across various scientific disciplines.

The primary contributions of this work are summarized as follows:

- We present PPN, a physics-inspired dual-branch framework specifically designed for ptychographic imaging that addresses the geometric mismatch between Euclidean spatial priors and concentric coherent patterns in reciprocal space.
- We propose the PoCA mechanism that replaces spatial neighborhood priors with radial-angular correlations, achieving superior high-frequency preservation compared to other end-to-end baselines.
- We demonstrate PPN's practical advantages with $>$ 1000$\times$ faster inference than iterative method, $<5\%$ performance degradation at 30% overlap ratio when tested on samples without feature distribution shifts.

II. Related Works

A. Deep Learning-Based Ptychographic Imaging

Deep learning in ptychographic imaging has recently enhanced computational efficiency and retrieval quality, categorized into three strategies:

1) Pre-Processing: Integrating deep learning with iterative algorithms improves initial estimates. The physics-informed automatic differentiation ptychography (ADP) framework uses pre-trained autoencoders to map high-dimensional image data to a low-dimensional latent space [15], while the double deep image prior (DDIP) method reduces the optimization parameter space [16], both enhancing convergence rates and noise robustness. However, these methods still require traditional iterative algorithms.

2) Post-Processing: Neural networks refine reconstructions from traditional algorithms, including enhancing a single iteration of the iterative algorithm to improve spatial resolution and reduce artifacts [5] [17]. While these methods significantly improve the quality of reconstructed images, they still rely on initial reconstructions and cannot provide real-time or fully automated solutions.

3) End-to-End with CNNs: End-to-end methods directly map diffraction patterns to complex object functions, bypassing iterative processes. Examples include PtychoNN with a modified U-Net and two-branch decoder for amplitude and phase [9], PtyNet with group convolution and Leaky ReLU for efficiency [11], and DPI using a traditional U-Net with skip connections [10]. Whereas CNNs excel at local feature extraction for natural images [18], [19], their inductive biases prove suboptimal for diffraction patterns requiring global phase coherence. The quadratic phase factors in Fresnel propagation create position-dependent correlations that span the entire detector plane, yet the local receptive fields of CNNs (typically 3×3) cannot span the full radial extent of diffraction rings, causing these physically critical phase relationships between distant pixels to be irrecoverably lost during feature encoding. Furthermore, progressive downsampling in hierarchical architectures systematically discards high-frequency information during feature abstraction. These inherent limitations of CNNs motivate the exploration of attention mechanisms for better long-range dependency modeling.

B. Transformer-Based Methods

Vision Transformers (ViTs) [20] emerged as a promising solution to the limitations of CNNs, demonstrating exceptional capability in capturing long-range dependencies across various domains [21], [22], [23], [24], [25]. Recent adaptations in ptychography, such as PtychoFormer [26] employs a hybrid architecture combining ViT and CNN components, specifically a SegFormer variant adapted for ptychography. While it surpasses traditional CNN methods in reconstruction accuracy, this comes with significant increases in parameter count and training costs. PtychoDV [17] adopts a compromise strategy that utilizes ViTs' reconstruction results as initial guesses for iterative methods, thereby enhancing the final accuracy of conventional iterative approaches. This indirectly reveals that pure end-to-end deep learning methods still face performance bottlenecks in diffraction reconstruction tasks - the core issue we aim to address in this paper.

While Transformers demonstrate superior capability in capturing global contexts, their direct application to diffraction patterns neglects crucial physical priors inherent in reciprocal space representations. Diffraction patterns, governed by wave optics, approximate the Fourier transform of an object's transmission function in far-field conditions [27], resulting in reciprocal space features like sparse representations and concentric rings [28]. Accurate image retrieval and phase retrieval require effectively capturing both sparsity and long-range dependencies within these patterns. Although Vision Transformers achieve global modeling through multi-head self-attention mechanisms (MHSA), their Euclidean spatial neighborhood-based attention weight calculation exhibits geometric mismatch with the concentric coherent patterns in reciprocal space characteristic of diffraction patterns. Our focus is to resolve the critical bottleneck of effective information extraction from diffraction patterns and improved reconstruction of high-frequency spatial information, rather than surpassing or replacing the ultimate resolution achieved by iterative algorithms.

III. Methodology

A. Problem Formulation

In ptychography, we aim to reconstruct a complex-valued object function $O(\mathbf{r})$, where $\mathbf{r}$ is the position vector in real space. This retrieval is based on a set of diffraction patterns measured at spatial frequencies $\mathbf{q}$ in reciprocal space. The process involves a probe function $P(\mathbf{r})$, which interacts with the object at various scanning positions $\{\mathbf{r}_j\}_{j=1}^{J}$. The ptychography problem can be formulated as:

$$\mathcal{F}\{\psi_j(\mathbf{r})\} = \sqrt{I_j(\mathbf{q})} \cdot \exp(i\phi_j(\mathbf{q})) \tag{1}$$

where $\psi_j(\mathbf{r}) = P(\mathbf{r}-\mathbf{r}_j)\cdot O(\mathbf{r})$ is the exit wave function, $\mathcal{F}$ denotes the Fourier transform, $I_j(\mathbf{q})$ is the measured diffraction intensity, and $\phi_j(\mathbf{q})$ is the phase of the diffraction pattern. To account for real-world factors, we model the diffraction intensity as:

$$\begin{aligned} I_j(\mathbf{q}) = \text{Poisson}\Big(\eta\Big|\mathcal{F}\{\mu P(\mathbf{r}-\mathbf{r}_j-\delta\mathbf{r}_j)O(\mathbf{r}) \\ +(1-\mu)P(\mathbf{r}-\mathbf{r}_j-\delta\mathbf{r}_j)\cdot\mathbb{E}_{\mathbf{r}}[O(\mathbf{r})]\}\Big|^2\Big) \\ +\mathcal{N}(0,\sigma^2) \end{aligned} \tag{2}$$

where μ is the coherence parameter, $\eta \in [0,1]$ is the detector efficiency representing the quantum efficiency (probability of detecting each photoelectron), Poisson(λ) denotes the Poisson distribution with rate parameter $\lambda = \eta\cdot|\mathcal{F}\{\dots\}|^2$, where $|\mathcal{F}\{\dots\}|^2$ is the theoretical diffraction intensity (in photons/pixel) and their product forms the expected value of the Poisson process, and $\delta\mathbf{r}_j$ accounts for positional jitter.

Based on the physical model of the actual imaging process in Eq. (2), we formulate the objective function for ptychographic retrieval in Eq. (3) to effectively recover the object and probe

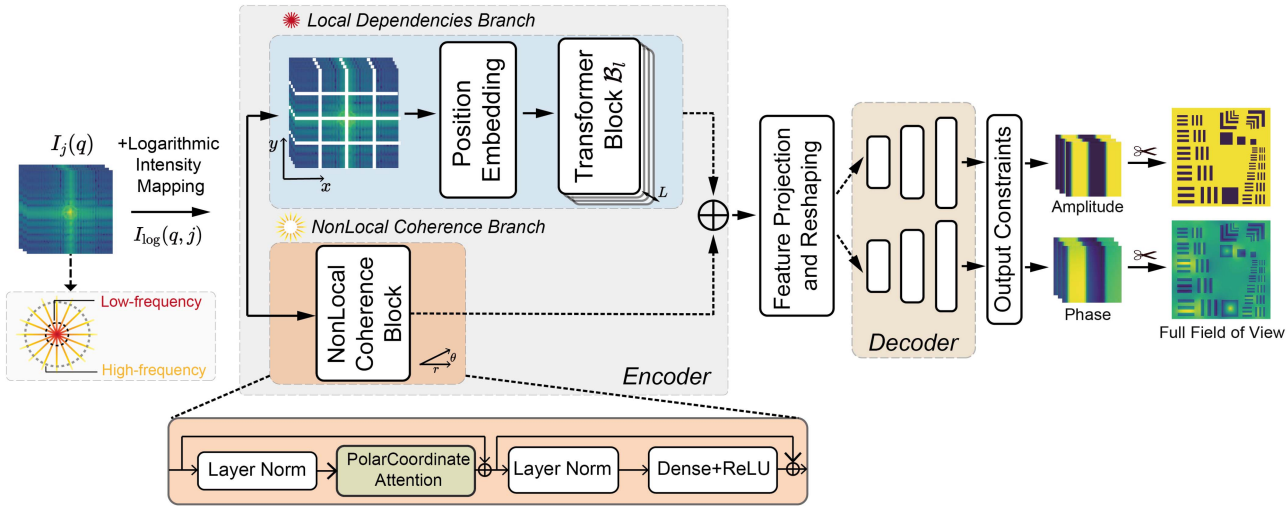

Fig. 2. The proposed PPN for ptychographic imaging. It features dual branches: a Local Dependencies Branch with standard ViT blocks, and a NonLocal Coherence Branch with Polar Coordinate Attention mechanism. The model processes logarithmically mapped diffraction patterns, combining features from both branches before decoding into individual amplitude and phase reconstructions at each position, which are then separately stitched to generate full field of view images for both amplitude and phase.

functions:

$$\mathcal{J}(O,P) = \frac{1}{J}\sum_{j=1}^{J} \left\| I_j - |\mathcal{F}\{P(\mathbf{r}-\mathbf{r}_j)\cdot O(\mathbf{r})\}|^2 \right\|_2^2 + \lambda_1\Omega_1(O) + \lambda_2\Omega_2(P) \tag{3}$$

where $\|\mathbf{x}\|_2^2 = \sum_i x_i^2$ is the squared Euclidean norm, $\Omega_1(O) = \|\nabla O\|_1$ enforces sparsity in the object gradient, $\Omega_2(P) = \|P - P_0\|_F^2$ constrains the probe function to maintain proximity to to an initial estimate P_0, and λ_1 and λ_2 are regularization weights.

B. PPN Architecture

An overview of the model architecture is provided in Fig. 2. The proposed architecture employs a bifurcated structure to map diffraction patterns to a reconstructed complex-valued object function. The input undergoes a logarithmic transformation $I_{\log}(\mathbf{q}, j) = \log(1 + I_j(\mathbf{q}))$ as a preprocessing step to address the large disparity between high and low frequency values in the diffraction patterns.

1) Local Dependencies Branch: The Local Dependencies Branch utilizes a standard Vision Transformer (ViT) to analyze diffraction patterns, capturing local features and structural relationships within each pattern. Operating on input tensors $\mathbf{X} \in \mathbb{R}^{B\times H\times W\times C}$ (where B is batch size, $H \times W$ is spatial resolution, and C is channel depth), this branch processes data through several key components:

Step 1. Patch Extraction divides the input image into non-overlapping patches of size $P \times P$, resulting in $\mathbf{X}_{\text{patches}} \in \mathbb{R}^{B\times N_p\times (P^2\cdot C)}$ where $N_p = \frac{H}{P} \times \frac{W}{P}$ represents the total number of patches.

Step 2. Linear Projection maps each patch to a D-dimensional embedding space using $\mathbf{X}_{\text{embed}} = \mathbf{X}_{\text{patches}}\mathbf{W}_E \in \mathbb{R}^{B\times N_p\times D}$ where $\mathbf{W}_E \in \mathbb{R}^{(P^2\,C)\times D}$ is the projection matrix.

Step 3. Positional Encoding adds spatial information via $\mathbf{X}_{\text{pos}} = \mathbf{X}_{\text{embed}} + \mathbf{E}_{\text{pos}}$, where the positional encoding matrix $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{N_p\times D}$ is broadcasted across the batch dimension to match $\mathbf{X}_{\text{embed}} \in \mathbb{R}^{B\times N_p\times D}$.

Step 4. Transformer Blocks with Pre-LN structure process the sequence through L layers according to $\mathbf{X}'^{(l)} = \mathcal{MSA}(\mathcal{LN}(\mathbf{X}^{(l-1)})) + \mathbf{X}^{(l-1)}$ and $\mathbf{X}^{(l)} = \mathcal{FFN}(\mathcal{LN}(\mathbf{X}'^{(l)})) + \mathbf{X}'^{(l)}$. The Multi-Head Self-Attention (MSA) mechanism computes $\mathcal{MSA}(\mathbf{X}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O$ where each attention head captures different relationship patterns in the data, with projection matrices $\mathbf{W}_i^Q, \mathbf{W}_i^K \in \mathbb{R}^{D\times d_k}$ and $\mathbf{W}_i^V \in \mathbb{R}^{D\times d_v}$ (typically $d_k = d_v = D/h$).

Step 5. Spatial Restoration reshapes the output to $\mathcal{H}_{\text{Local}} \in \mathbb{R}^{B\times(H/P)\times(W/P)\times D}$.

The complete dimension flow is: $\mathbb{R}^{B\times H\times W\times C} \xrightarrow{(1)} \mathbb{R}^{B\times N_p\times(P^2\,C)} \xrightarrow{(2)} \mathbb{R}^{B\times N_p\times D} \xrightarrow{(3)} \mathbb{R}^{B\times N_p\times D} \xrightarrow{(4)} \mathbb{R}^{B\times N_p\times D} \xrightarrow{(5)} \mathbb{R}^{B\times(H/P)\times(W/P)\times D}$.

2) Non-Local Coherence Branch: The Non-Local Coherence Branch employs an architectural framework similar to the Local Dependencies Branch, sharing identical transformer block components while introducing our novel Polar Coordinate Attention (PoCA) mechanism. This specialized attention mechanism processes input at pixel-level granularity ($H \times W$ pixels) without patch extraction, enabling the capture of long-range dependencies across diffraction patterns. The PoCA mechanism systematically encodes physical constraints inherent to diffraction physics, including the $I \propto r^{-4}$ intensity distribution through radial decay weighting, preserves crystallographic symmetry via angular continuity, and adapts to illumination shifts with dynamically learned centering parameters α_x and α_y.

After logarithmic transformation, the single-channel diffraction pattern $\mathbb{R}^{B\times H\times W\times 1}$ is projected to a multi-channel tensor $\mathbb{R}^{B\times H\times W\times C}$ via a learnable 1×1 convolutional layer, where C denotes the embedding dimension. For an input tensor $\mathbf{X} \in \mathbb{R}^{B\times H\times W\times C}$ (where B represents batch size, $H \times W$ spatial resolution), this branch processes data through several key components:

Step 1. Tensor Flattening transforms the input into a sequence representation via $\mathbf{X}_f = \text{Reshape}(\mathbf{X}, (B, H \times W, C)) \in \mathbb{R}^{B\times N\times C}$, where $N = H \times W$ represents the total number of pixels.

Step 2. Query-Key-Value Projection maps the flattened tensor into three separate embedding spaces using:

$$\begin{aligned} \mathcal{Q}^i &= \mathbf{X}_f \mathbf{W}_Q^i \qquad \mathbf{W}_Q^i \in \mathbb{R}^{C\times d_k} \\ \mathcal{K}^i &= \mathbf{X}_f \mathbf{W}_K^i \qquad \mathbf{W}_K^i \in \mathbb{R}^{C\times d_k} \\ \mathcal{V}^i &= \mathbf{X}_f \mathbf{W}_V^i \qquad \mathbf{W}_V^i \in \mathbb{R}^{C\times d_v} \end{aligned} \tag{4}$$

where projection matrices $\mathbf{W}_Q^i, \mathbf{W}_K^i \in \mathbb{R}^{C\times d_k}$ and $\mathbf{W}_V^i \in \mathbb{R}^{C\times d_v}$ define the embedding transformations for attention head i. The resulting tensors are $\mathcal{Q}^i, \mathcal{K}^i \in \mathbb{R}^{B\times N\times d_k}$ and $\mathcal{V}^i \in \mathbb{R}^{B\times N\times d_v}$, where the matrix multiplication is performed along the feature dimension C. Polar coordinate parameterization defines the physical diffraction geometry using:

$$c_x = \frac{W}{2} + \alpha_x \frac{W}{2}, \quad c_y = \frac{H}{2} + \alpha_y \frac{H}{2} \tag{5}$$

$$\begin{aligned} r_m &= \frac{\log(1 + \|\mathbf{p}_m - \mathbf{c}\|)}{\log(1 + r_{\max})}, \quad \theta_m \\ &= \arctan 2\,(y_m - c_y, x_m - c_x) \end{aligned} \tag{6}$$

$$\mathbf{\Phi}_r^{mn} = \frac{1}{1 + |r_m - r_n|}, \quad \mathbf{\Phi}_\theta^{mn} = \cos(\theta_m - \theta_n) \tag{7}$$

where $r_{\max} = \sqrt{(W/2)^2 + (H/2)^2}$ ensures normalized radial coordinates, $\mathbf{p}_m = (x_m, y_m)$ denotes original pixel coordinates, $\mathbf{c} = (c_x, c_y)$ represents the learned diffraction center, and $\alpha_x, \alpha_y \in [-0.5, 0.5]$ are learnable parameters initialized at 0. The implementation employs logarithmic scaling to compress the dynamic range of radial distances, with angle normalization $\theta_m \in [0, 2\pi)$ ensuring continuity.

Step 3. Polar Coordinate Modulation incorporates physical constraints by modulating the standard dot-product attention with radial and angular weighting matrices:

$$\mathcal{A}_{\text{polar}}^i = \underbrace{\frac{\mathcal{Q}^i (\mathcal{K}^i)^\top}{\sqrt{d_k}}}_{\text{base attention}} \odot \underbrace{\mathbf{\Phi}_r}_{\text{radial decay}} \odot \underbrace{\mathbf{\Phi}_\theta}_{\text{angular continuity}} \tag{8}$$

where $\mathbf{\Phi}_r$, $\mathbf{\Phi}_\theta \in \mathbb{R}^{1\times N\times N}$ (broadcastable to $\mathbb{R}^{B\times N\times N}$), $\odot$ represents element-wise multiplication with broadcasting, imposing both radial decay and angular continuity constraints.

Step 4. Attention Application computes each attention head through $\text{Head}_i = \text{softmax}(\mathcal{A}_{\text{polar}}^i)\mathcal{V}^i$, where the softmax normalization ensures proper probability distribution across the attention weights, and $\mathcal{A}_{\text{polar}}^i \in \mathbb{R}^{B\times N\times N}$ is applied to $\mathcal{V}^i \in \mathbb{R}^{B\times N\times d_v}$.

Step 5. Multi-head Integration combines information from all h attention heads via $\mathcal{G}(\mathbf{X}_f) = \text{Concat}(\text{Head}_1, \ldots, \text{Head}_h)\mathbf{W}^O$, with $\mathbf{W}^O \in \mathbb{R}^{h\cdot d_v\times C'}$ serving as the output projection matrix and $\text{Head}_i \in \mathbb{R}^{B\times N\times d_v}$.

Step 6. Spatial Restoration reshapes the processed output to match the original spatial dimensions through $\mathcal{P}oCA(\mathbf{X}) = \text{Reshape}(\mathcal{G}(\mathbf{X}_f), (B, H, W, C')) \in \mathbb{R}^{B\times H\times W\times C'}$.

The complete dimension flow through this branch is: $\mathbb{R}^{B\times H\times W\times C} \xrightarrow{(1)} \mathbb{R}^{B\times N\times C} \xrightarrow{(2)} \{\mathbb{R}^{B\times N\times d_k} \quad (\mathcal{Q}^i, \mathcal{K}^i), \mathbb{R}^{B\times N\times d_v} (\mathcal{V}^i)\} \xrightarrow{(3)} \mathbb{R}^{B\times N\times N} \xrightarrow{\text{softmax}} \mathbb{R}^{B\times N\times N} \xrightarrow{(4)} \mathbb{R}^{B\times N\times d_v} \xrightarrow{(5)} \mathbb{R}^{B\times N\times h\cdot d_v} \xrightarrow{\mathbf{W}^O} \mathbb{R}^{B\times N\times C'} \xrightarrow{(6)} \mathbb{R}^{B\times H\times W\times C'}$

By decomposing attention into radial and angular components, PoCA effectively mimics the Ewald sphere's intersection with the reciprocal lattice (Fig. 1(b)), where r_{mn} correlates with the reciprocal space resolution of scattering vectors, and θ_{mn} encodes Bragg angle relationships through angular coherence constraints.

3) Feature Fusion and Decoding: The feature fusion process addresses the resolution mismatch between the two branches. The Local Dependencies Branch output $\mathcal{H}_{\text{Local}} \in \mathbb{R}^{B\times(H/P)\times(W/P)\times D}$ has lower spatial resolution than the Non-Local Branch output $\mathcal{H}_{\text{NonLocal}} \in \mathbb{R}^{B\times H\times W\times C'}$. To align these representations, we employ bilinear upsampling on the Local Branch features:

$$\mathcal{H}_{\text{Local}}^{\text{upsampled}} = \text{Upsample}_{\text{bilinear}}(\mathcal{H}_{\text{Local}}) \in \mathbb{R}^{B\times H\times W\times D} \tag{9}$$

The aligned features are then combined through concatenation followed by a $1\times$ 1 convolution:

$$\mathcal{H}_{\text{fused}} = \text{Conv}_{1\times 1}(\text{Concat}[\mathcal{H}_{\text{Local}}^{\text{upsampled}}, \mathcal{H}_{\text{NonLocal}}]) \tag{10}$$

The decoder $\mathcal{D}$ consists of three transposed convolution blocks with progressive upsampling:

$$\begin{aligned} \mathcal{D}_i &= \text{BatchNorm}(\text{ReLU}(\text{ConvTranspose2D}_{k,s})) \\ \mathcal{D} &= \mathcal{D}_3 \circ \mathcal{D}_2 \circ \mathcal{D}_1 \end{aligned} \tag{11}$$

where k is kernel size and s is stride. The decoder branches into parallel paths $\mathcal{D}_{\text{amplitude}}$ and $\mathcal{D}_{\text{phase}}$ to reconstruct the object's amplitude and phase components.

4) Training Objective: The loss function $\mathcal{L} : \Theta \to \mathbb{R}^+$ is defined as:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{n=1}^{N} \left[\left\| A_{\text{true}}^{(n)} - A_{\text{pred}}^{(n)} \right\|_2^2 + \left\| \phi_{\text{true}}^{(n)} - \phi_{\text{pred}}^{(n)} \right\|_2^2 \right] \tag{12}$$

where N is the number of training samples, $A_{\text{true}}^{(n)}$ and $\phi_{\text{true}}^{(n)}$ are the true amplitude and phase for the n-th sample (both amplitude and phase values are normalized to the range [0,1]), respectively, and $A_{\text{pred}}^{(n)} = A_{\text{final}}^{(n)}$ and $\phi_{\text{pred}}^{(n)} = \phi_{\text{final}}^{(n)}$ are the corresponding model predictions. The two error terms naturally maintain comparable scales without requiring additional weighting factors.The optimization problem is formulated as:

$$\theta^* = \arg\min_{\theta\in\Theta} \mathcal{L}(\theta) \tag{13}$$

where θ^* represents the optimal model parameters that minimize the loss function $\mathcal{L}(\theta)$ over the training dataset.

IV. Experiments and Results

A. Experiment Setup

1) Implementation Details: We implemented both our DL-based method and the traditional iterative algorithm (like the extended ptychographic iterative engine (ePIE) algorithm [29]) in Python with CUDA acceleration on the same hardware platform. The iterative algorithm was implemented with CUDA support, initialized with known probe measurements and processing 32× 32 diffraction patterns, with phase unwrapping applied to retrieve the continuous phase. The ePIE implementation typically required 100-200 iterations to converge and 75-80% spatial overlap, with minimal gains beyond. Both implementations were trained and evaluated on a NVIDIA Tesla T4 GPU with Intel Xeon Platinum 8480CL CPU @ 2.00GHz running Linux 6.1.85+. For the DL-based method, we used TensorFlow with an MSE loss function and the Adam optimizer. We set the initial learning rate to 1.0 and implemented a ReduceLROnPlateau callback (factor 0.5, patience 2, $\min_{\text{lr}}$ 0.0001) to dynamically adjust the learning rate based on validation loss. We used a batch size of 32 with 5% validation split, while a custom callback monitored system memory usage, computation time, and convergence behavior.

2) Datasets: We utilize two types of datasets in our experiments: **Simulated Dataset**: The dataset consists of a 512× 512 LTEM image [30]. The pattern is based on a slightly modified version of the 1951 USAF resolution test chart. Using a probe size of 32× 32 pixels with a 75% overlap rate and 3-pixel jitter, we generated a total of 3721 patches arranged in a 61× 61 grid, each patch being 32× 32 pixels. Gaussian noise simulates realism, with added Poisson and Gaussian noise for readout and photon effects. Amplitude and phase undergo minmax normalization. **Real Experimental Data**: Experimental data from Argonne National Laboratory [9] comprises 16,100 triplets of diffraction data, amplitude, and phase images. 161 × 161 point scan at 30 nm increments from X-ray nanoprobe beamline 26-ID. PIE algorithm generates ground truth. For both datasets, we split the data into 80% for training and 20% for testing. The validation set is created using 5% of the training data. Results on simulated data are presented in Sections IV.B, D, E, F, G and Section V.B, while Section IV.C validates the model using real experimental synchrotron data.

3) Evaluation Metrics: The performance of the models is evaluated using MSE, Peak Signal-to-Noise Ratio (PSNR), and SSIM, with the ground truth values of the materials basis images as references.

B. Performance Comparison and Analysis

We evaluated PPN against three mainstreamed CNN-based methods (PtychoNN, PtyNet, and DPI). Our analysis covered single-shot retrieval and full-stitched field retrieval under both ideal and noisy conditions.

1) Single Scan Point Retrieval Analysis: PPN reconstructed retrievals with improved edge definition and clarity compared to those from the other methods in single scan point recovery, where each position represents an individual amplitude/phase reconstruction from a corresponding diffraction pattern. This improvement is particularly evident in the vertical edge on the left side of Fig. 3(a). The reconstructed images also showed contrast levels closer to the ground truth, with more natural transitions between dark and bright areas. By contrast, the CNN-based models tended to smooth these sharp features and introducing artifacts and distortions accuracy.

2) Full-Stitched Field Retrieval Analysis: Table I presents full-field stitching results under ideal (noise-free) and realistic (noisy) conditions, demonstrating PPN's superior performance across all metrics. In ideal conditions, PPN achieved 16.0% lower MSE_{amp} and 6.7% higher PSNR_{amp} compared to PtychoNN, with similar improvements under noisy conditions. Fig. 3(c) visually confirms PPN's superior global consistency and minimal boundary artifacts, particularly in reconstructing vertical column ends and background textures. The dual-branch structure enables our model to effectively capture both low-frequency and mid-to-high frequency information of diffraction patterns. This capability stems from the model's ability to capture the intrinsic geometric structure of the data, analogous to continuous mapping on high-dimensional manifolds.

Repeated measures ANOVA and pairwise t-tests (Bonferroni correction, $\alpha = 0.0083$) showed PPN significantly outperformed the other methods in all metrics in both conditions ($p < 0.0083$), especially in MSE_{amp} and SSIM_{amp} ($p < 0.0001$). The three CNN-based methods showed no significant differences, performing similarly across most metrics. Fig. 3 focuses on intra-family comparisons among deep learning architectures to isolate the impact of structural variations under consistent training supervision. Comparative evaluations with conventional iterative algorithms (e.g., ePIE) are presented separately in Section IV-E3, where differences in computational efficiency and resilience to varying overlap ratios are quantitatively assessed and visually demonstrated.

3) Spatial Frequency Fidelity Analysis: We analyze frequency-domain fidelity using 1D diagonal cross-sections of the averaged power spectral density (PSD), defining three bands by normalized radial distance: low (0-1/3 Nyquist), mid (1/3-2/3), and high (2/3-1) frequencies.

Fig. 4(a) reveals PPN closely matches the ground truth's PSD profile, particularly preserving mid-frequency features (red box). CNN-based methods show $2.1-3.7\times$ greater mid-frequency attenuation (0.06–0.07% vs. 0.22% total energy) and high-frequency suppression ($< 0.03\%$). Quantitatively, PPN achieves 58.6% better mid-frequency preservation than CNNs while maintaining noise-equivalent high-frequency levels (0.03% vs. ground truth's 0.04%). This frequency-selective enhancement stems from our polar coordinate attention mechanism, which preferentially weights mid-frequency correlations corresponding to Bragg diffraction conditions while suppressing high-frequency noise through radial decay constraints. The results confirm PPN's unique ability to resolve genuine high-resolution features among the end-to-end methods.

C. Validation on Experimental Synchrotron Data

Fig. 5 demonstrates PPN's superior performance on real experimental data, consistent with our findings from simulated datasets. Across all metrics (PSNR, SSIM, and MSE) for both

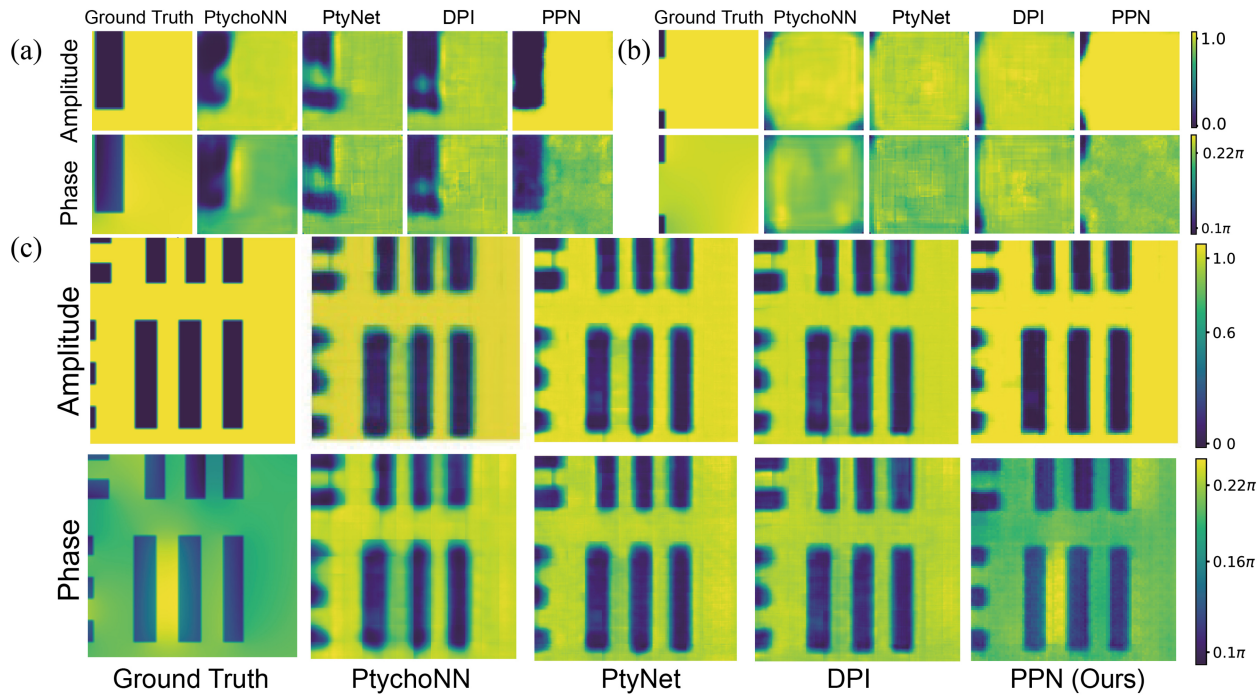

Fig. 3. Performance comparison of single-shot experiment results and full-stitched scene retrieval using simulated data. (a) and (b) show amplitude and phase reconstructions at two representative scan positions. (c) displays the full-field amplitude and phase images stitched together from individual reconstructions at all scan positions across different models.

TABLE I
QUANTITATIVE COMPARISON OF FULL-STITCHED SCENES UNDER NOISE-FREE AND NOISY CONDITIONS ON SIMULATED DATA

Condition	Method	Amplitude			Phase		
		MSE ($\times 10^{-2}$) ↓	PSNR (dB) ↑	SSIM (%) ↑	MSE ($\times 10^{-2}$) ↓	PSNR (dB) ↑	SSIM (%) ↑
Noise-free	PtychoNN [9]	4.32 ± 0.02	13.52 ± 0.17	83.50 ± 0.80	1.22 ± 0.01	19.09 ± 0.19	67.10 ± 0.70
	PtyNet [11]	4.46 ± 0.03	13.56 ± 0.15	83.10 ± 0.60	1.28 ± 0.01	19.10 ± 0.26	67.80 ± 1.30
	DPI [10]	4.29 ± 0.03	13.68 ± 0.12	83.40 ± 0.50	1.26 ± 0.01	18.91 ± 0.24	67.70 ± 1.10
	PPN (Ours)	**3.63 ± 0.04**	**14.42 ± 0.08**	**87.00 ± 0.30**	**1.14 ± 0.01**	**19.41 ± 0.05**	**70.60 ± 0.70**
Noisy	PtychoNN	6.66 ± 0.03	11.76 ± 0.12	74.00 ± 0.89	1.47 ± 0.04	18.33 ± 0.17	58.90 ± 1.00
	PtyNet	6.50 ± 0.02	11.91 ± 0.13	74.80 ± 0.66	1.43 ± 0.07	18.46 ± 0.18	62.10 ± 1.50
	DPI	6.46 ± 0.02	11.76 ± 0.10	74.20 ± 0.65	1.48 ± 0.06	18.31 ± 0.20	62.00 ± 1.20
	PPN (Ours)	**5.91 ± 0.03**	**12.34 ± 0.11**	**78.80 ± 0.59**	**1.31 ± 0.06**	**18.76 ± 0.16**	**68.30 ± 1.10**

Note: For noisy conditions, noise is simulated using Gaussian and Poisson distributions to model readout noise and photon noise, respectively. Bold values indicate the best performance for each metric. PSNR in dB, and SSIM in percentage. ↑: higher is better, ↓: lower is better. Results are presented as mean ± standard deviation from ten independent experiments.

amplitude and phase retrieval, PPN consistently outperforms other models. These improvements are visible in fine structural details (Fig. 5(a), (b)). The model's ability to preserve edge sharpness and resolve intricate features in experimental data validates its potential for practical applications where accurate reconstruction of complex nanostructures is essential for scientific interpretation.

D. Generalization Capabilities Comparison

We evaluated PPN's generalization by training on simulated 1951 USAF pattern (the same with Section IV-B) with straight-line features and evaluating on complex, curved patterns (Fig. 6), without fine-tuning, assessing the model's ability to extrapolate learned features to significantly different sample geometries. This evaluation connects directly to practical applications where pre-trained models must process new samples with unknown structures—a scenario where acquiring specific training data is often impractical. In these contexts, generalization capability determines the practical utility of deep learning for ptychographic imaging. Fig. 6 compares phase retrieval results across different methods. While deep learning models cannot match the reconstruction quality of the ePIE algorithm, PPN outperforms all other DL-based methods. The red boxes highlight high-frequency details where PPN preserves spatial information with minimal artifacts compared to other end-to-end methods.

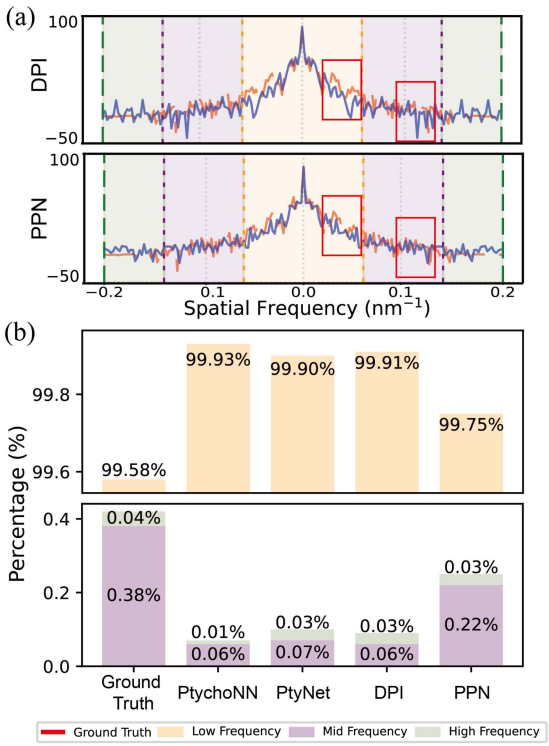

Fig. 4. Frequency analysis comparison of full-scene ptychographic retrievals on simulated data. (a) 1D diagonal cross-sections of average 2D PSD from stitched simulations. Red curve: ground truth; Blue curve: models. The red box represents obvious abnormal model retrieval. (b) Quantitative breakdown of energy distribution across low, mid, and high frequency bands. Frequency ranges are defined based on the radial distance from the PSD center, with boundaries at 1/3 and 2/3 of the maximum frequency.

Results show that even with different feature distributions between training and testing data, PPN achieves superior reconstruction quality and better preservation of high-frequency spatial details.

E. Performance Comparison Under Limited Data

Despite significant advancements in ptychography, a major challenge remains: extended data acquisition periods. While higher overlap between scanning points typically yields better retrieval results [14], it also increases experiment duration and radiation exposure. This is particularly problematic for radiation-sensitive materials [31] and in *situ* dynamic studies [32]. Here, the overlap ratio is defined as the physical overlap between adjacent scanning positions in real space.

1) Comparison With Different Training Data Size: We evaluated the effect of training data size on model performance by training models with 10 different random initializations while keeping all hyperparameters consistent and testing on the same test set as in Section IV-B. As shown in Fig. 7(a), when the training samples decrease from 2928 to 50, PPN demonstrates superior data efficiency: the PSNR for amplitude reconstruction decreases by only 35%, and phase reconstruction by 15%. In contrast, PtychoNN shows the most significant performance degradation, with PSNR dropping by approximately 70% for both amplitude and phase reconstruction. This resilience, analyzable through compressed sensing theory. [33], suggests PPN learns an optimal sparse prior.

2) Comparison Across Overlap Ratios With a Fixed Training Size: We investigated the impact of test set overlap ratios (0%, 25%, 50%, and 75%) while maintaining a fixed training set size of 100 samples, consistent with the models used in Section IV-E1. While the test area maintains consistent spatial coverage (matching Fig. 3(c)'s field of view), the number of test samples decreases with lower overlap ratios due to reduced spatial sampling density. As illustrated in Fig. 7(b), all models exhibit remarkable metric stability across different overlap ratios. Taking PPN as an example, its amplitude reconstruction PSNR remains stable between 9.5–9.6 ($\Delta = 0.1$) and SSIM between 0.39–0.40 ($\Delta = 0.01$), with one-way ANOVA ($p > 0.05$) confirming no significant differences between overlap ratio groups.

This overlap-invariant behavior stems from our training approach, which uses simulated real-space amplitude and phase as ground truth—data that typically requires exceptionally high overlap ratios to obtain in conventional experimental workflows. This characteristic enables an optimal workflow: using high-overlap data for model training while deploying at low overlap during testing. These findings naturally lead to the question of how PPN compares to traditional iterative methods across different overlap ratios, which we explore in the following section.

3) Comparison With Iterative Methods Across Overlap Ratios: We compared PPN with the iterative method (where we use the widely-used ePIE as our baseline) across various overlap ratios to assess both reconstruction quality and computational efficiency. Fig. 8 presents this comparison, with PPN trained using the same dataset described in Section IV-B, where sample features remain within the learned distribution. As shown in Fig. 8(a), PPN maintains SSIM values between 0.86–0.92 across all overlap ratios, indicating consistent reconstruction quality. In contrast, iterative method exhibits strong overlap dependence, with SSIM dropping from 0.95 at 90% overlap to only 0.12 at 30% overlap. At 30% overlap, PPN completes reconstructions in approximately 0.15 seconds versus iterative method's 125 seconds, representing an 852× speed improvement, and at 60% overlap, a 1767× speed improvement. The reconstruction time was measured by averaging over ten consecutive tests, with error bars representing the variance. Reducing overlap from 90% to 30% increases data acquisition efficiency by approximately 49 times for a fixed area, as scan point density is inversely proportional to the square of the step size. Theoretically, this combined effect could yield up to 41,748× (49 × 852) overall efficiency improvement. This enables significantly faster experiments with substantially reduced radiation exposure while still producing high-quality reconstructions.

This separation of training and testing strategies, along with PPN's efficiency, demonstrates the synergy between deep learning and domain-specific knowledge in ptychography, providing a viable technical pathway for achieving high-throughput imaging.

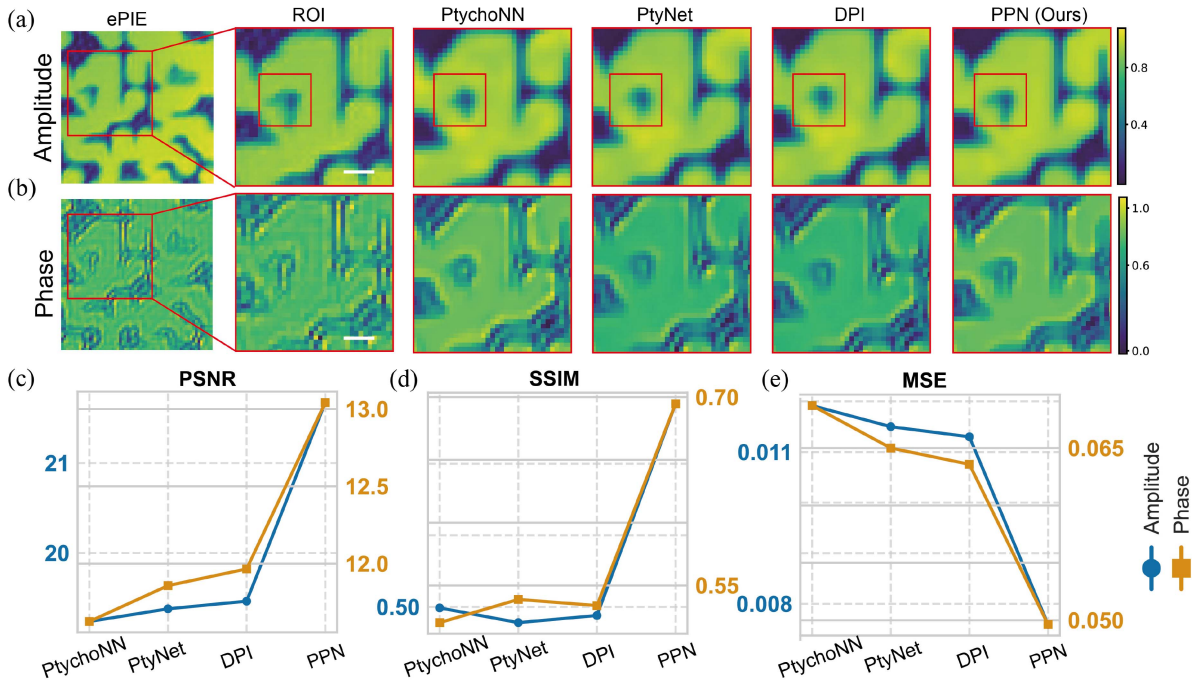

Fig. 5. Performance comparison on real experimental samples. (a,b) Visual comparison of retrieved amplitude and phase (scale bar: 200 nm). For a specific hook-shaped detail in ROI (region of interest), only our model effectively restores it, significantly outperforming CNN-based methods in fine structure retrieval. (c-e) Quantitative comparison of PSNR, SSIM, and MSE.

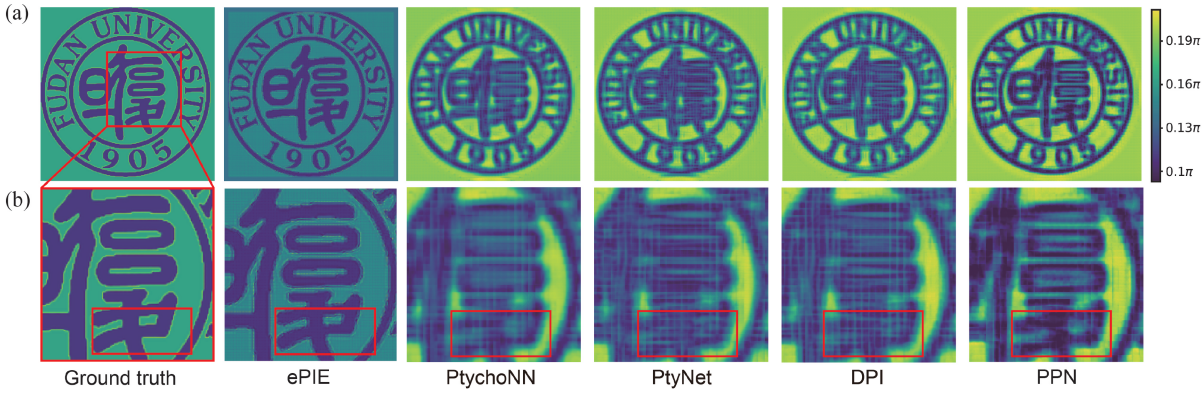

Fig. 6. Comparison of generalization capabilities across different methods using the Fudan University logo (simulated data) as a test sample with distribution significantly different from the training set. (a) Phase retrieval results with intensity values normalized to π. (b) Detailed ROI comparison.

F. Ablation Study

To justify each component of the proposed PPN, we conducted comprehensive ablation studies. As shown in Fig. 9, we compared both amplitude and phase reconstructions, including 2D reconstructed images and 1D cross-sectional profiles. We examined several model variants: 1) Ours without NonLocal Coherence Branch (Ours w/o NLB), which means using only the Local Dependencies Branch and removing the NonLocal Coherence Branch that contains proposed PoCA; 2) Ours with Multi-Head Self-Attention (MHSA) in NLB (Ours w/ MinN), replacing PoCA with standard MHSA in the NonLocal Branch; 3) Ours without Decoder (Ours w/o D), using fully connected layers instead of CNN for upsampling in the decoder; and 4) Ours (Full Model), the full proposed model. The ablation experiments were conducted using the same training set, test set, and parameter settings as in Section IV-B.

The quantitative results are presented in Table II, where bold values indicate the best performance for each metric. The removal of the NonLocal Coherence Branch leads to decreased structural coherence in phase reconstruction, which is clearly observable in the 1D profiles. Replacing PoCA with standard MHSA results in a 4.16% drop in PSNR_{amp} and 3.56% in SSIM_{amp}, with poorer edge transitions visible in the profile plots. The absence of the CNN decoder causes the most significant performance degradation, with PSNR_{amp} and

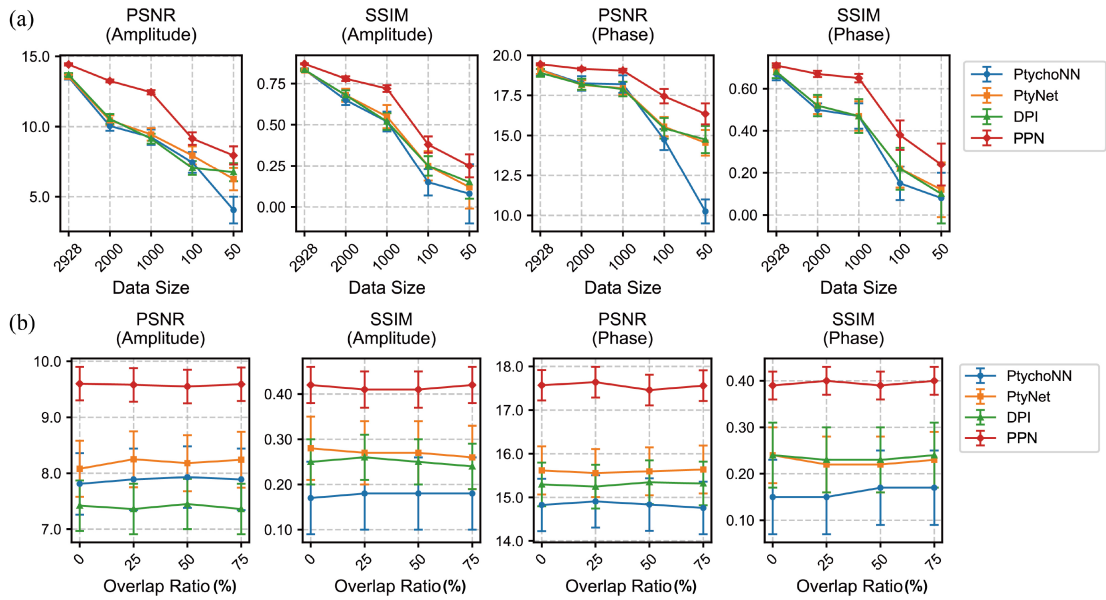

Fig. 7. Model performance evaluation on simulated data. (a) Reconstruction quality versus training data size at overlap ratio 75%. (b) Performance stability under different test set overlap ratios with fixed training set size (100 samples).

TABLE II
ABLATION STUDY RESULTS ON SIMULATED DATA

Model Variant	Amplitude		Phase	
	PSNR (dB)	SSIM (%)	PSNR (dB)	SSIM (%)
Ours w/o NLB	14.10	86.60	18.39	65.40
Ours w/ MinN	13.82	83.90	18.64	64.10
Ours w/o D	11.02	60.50	17.32	54.50
Ours	**14.42**	**87.00**	**19.41**	**70.60**

Bold value indicate the best performance for each metric.

TABLE III
PERFORMANCE COMPARISON OF MODEL VARIANTS

Method	Amplitude		Phase	
	PSNR (dB)	SSIM (%)	PSNR (dB)	SSIM (%)
TransUNet [34]	13.57	81.20	18.77	67.80
SegFormer [35]	13.76	83.30	19.12	68.10
PtychoFormer [26]	13.77	83.54	19.17	68.18
PPN (Ours)	**14.42**	**87.00**	**19.41**	**70.60**

Bold value indicate the best performance for each metric.

SSIM_{amp} decreasing by 23.58% and 30.46%, respectively. The reconstructed images without the CNN decoder appear blurry, consistent with the experimental results in [17]. The 1D profiles reveal that without the CNN decoder, the model struggles to maintain accurate amplitude levels and phase transitions. The full model achieves the best performance across all metrics, particularly in regions with dramatic phase changes in the 1D cross-sectional profiles. These results validate the effectiveness of each proposed component in our architecture.

G. Model Variants

To validate our model's effectiveness, we benchmarked against two widely-adopted hybrid CNN-ViT architectures from computer vision: TransUNet [34] and SegFormer [35]. These baselines remain relevant in recent research across multiple domains [36], [37], [38], [39], demonstrating their continued utility for important results in various fields.

The first variant adopts a TransUNet-inspired architecture [34], employing a hybrid CNN-Transformer encoder where CNNs extract initial features before Transformer processing, with a decoder utilizing a cascading structure and skip connections. The second implements a SegFormer-based approach [35], featuring a hierarchical structure with progressive resolution reduction for multi-scale feature extraction and replacing the first linear layer in the feed-forward network with a 3×3 convolution, similar to the approach used in PtychoFormer [26]. PtychoFormer uses an encoder based on SegFormer and a decoder adapted specifically for ptychography tasks. Experiments conducted under noise-free conditions (see Section IV-B2) demonstrate that PPN outperforms all baseline models (Table III). The PtychoFormer results are comparable to our SegFormer-inspired model due to their similar structures. These variants' limitations stem from designs optimized for real-space image processing: The TransUNet variant, while benefiting from the combination of CNN and Transformer, still relies on local feature extraction in its initial stages. This approach is suboptimal for capturing the global coherence information present in diffraction patterns. Similarly, the SegFormer variant's hierarchical structure, while effective for multi-scale feature extraction in natural images, may lose critical high-frequency information in the context of diffraction patterns due to its progressive downsampling.

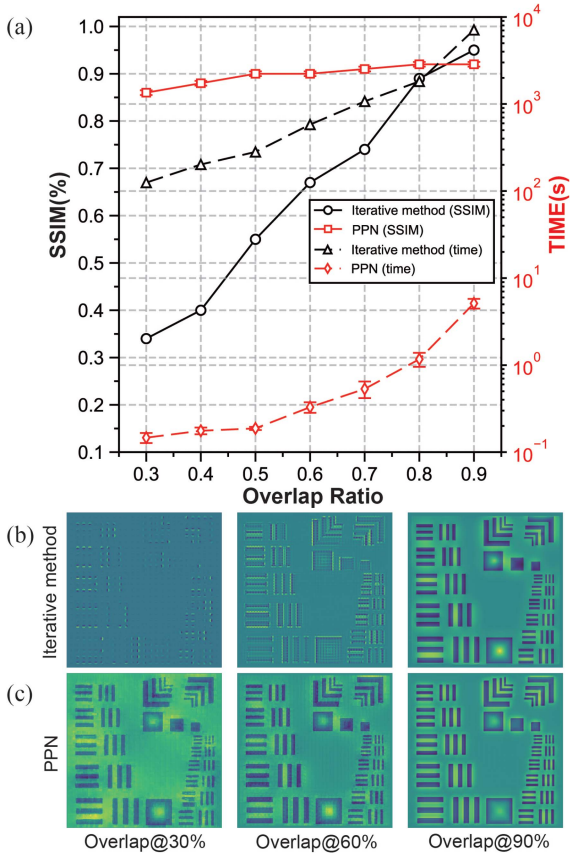

Fig. 8. Comparison between iterative method and PPN across different overlap ratios. (a) SSIM and reconstruction time comparison. (b) and (c) Reconstructed images using iterative method and our proposed PPN, respectively, at different overlap ratios (30%, 60%, 90%).

V. Discussion

A. Computational Complexity Analysis

Our analysis compares PPN with CNN-based methods (PtychoNN, PtyNet, DPI) and the ViT-CNN hybrid baseline (PtychoFormer [26]) through computational metrics in Fig. 10. While maintaining hierarchical CNN decoders common to all compared methods, our encoder design achieves substantial efficiency gains.

Among CNN approaches, PPN demonstrates intermediate complexity (6.12 MB parameters) between PtyNet (1.24 MB) and DPI (57.14 MB), yet converges faster (30 epochs) than both PtychoNN (45 epochs) and DPI (35 epochs). The 0.382 GFLOPs cost represents a 63.7% reduction from DPI's 1.049 GFLOPs, confirming efficient feature learning despite transformer integration. Compared to PtychoFormer's ViT-CNN architecture, our model reduces parameters by 91.1% (6.12 MB vs. 68.96 MB) and FLOPs by 65.4% (0.382 G vs. 1.105 G) while matching convergence speed (30 epochs). This efficiency comes from three optimizations: 1) fixed feature dimensions preventing channel inflation, 2) parallel dual-branch processing instead of hierarchical refinement, and 3) physics-inspired attention limiting learnable parameters. These changes eliminate redundant operations while preserving physical priors.

TABLE IV
Comparison With Different Loss Functions

Loss Function	Amplitude		Phase	
	PSNR (dB)	SSIM (%)	PSNR (dB)	SSIM (%)
MSE Loss	14.42	87.00	19.47	70.60
MAE Loss	13.63	85.90	19.29	67.50
Huber Loss	14.64	87.40	19.54	68.40
NPCC Loss	10.36	69.90	17.03	66.90
NSSIM Loss	14.12	85.10	19.41	70.80
Weighted Loss ($\alpha = 0.9$)	**15.10**	**88.50**	**20.04**	**73.90**

Note:The weighted loss combines MSE and NSSIM losses, where α represents the weight of MSE (0.9) and 1-α (0.1) is the weight of NSSIM. Bold value indicate the best performance for each metric.

B. Loss Function Analysis

To optimize retrieval quality in ptychographic phase retrieval, we evaluated various loss functions (as shown in Table IV. Standard metrics like MSE and Mean Absolute Error (MAE) showed similar performance, while the Huber loss [40], defined as: $L_{\text{Huber}}(x) = \begin{cases} \frac{1}{2}x^2 & \text{for } |x| \le \delta \\ \delta(|x| - \frac{1}{2}\delta) & \text{otherwise} \end{cases}$ offered robustness to outliers. The Negative Pearson Correlation Coefficient (NPCC) loss, despite its success in single-frame phase retrieval [12], showed poor convergence stability in our experiments. This is likely due to its sensitivity to local statistics in small ptychographic patches, contrasting with its effectiveness on larger, more statistically stable single-frame images.

We introduced a novel combined loss function inspired by the work on image restoration with neural networks [42]: $L_{\text{combined}}(\alpha) = \alpha L_{\text{MSE}} + (1-\alpha)L_{\text{NSSIM}}$ where $L_{\text{NSSIM}} = -\text{SSIM}(I_q, |\mathcal{F}P(r)\psi(r)|^2)$, where $\psi(r)$ defined as $\psi(r) = P(r - r_j)O(r)$. This approach balances global consistency (MSE) with local structural preservation (Negative Structural Similarity Index Measure (NSSIM) [43]). Experiments revealed optimal performance at $\alpha = 0.9$, with consistent improvement as α increased from 0.2 to 0.9, as shown in Fig. 11. This robustness to α values reduces the need for precise tuning in practical applications. The combined loss function excels in addressing affine ambiguity, as ∇L_{MSE} is sensitive to global affine transformations T ($\frac{\partial L_{\text{MSE}}}{\partial T} \neq 0$), while ∇L_{NSSIM} maintains local structure invariance ($\frac{\partial L_{\text{NSSIM}}}{\partial(\text{local structure})} \approx 0$), forming a more robust optimization target.

VI. Conclusion

We propose PPN, a physics-inspired deep learning framework that significantly improves ptychographic reconstruction through two core innovations: (1) an architecture combining local feature extraction with global pattern coherence modeling, and (2) a novel attention mechanism tailored for diffraction physics. PPN demonstrated superior performance over existing state-of-the-art methods, particularly in high-frequency artifact suppression and data efficiency. Current limitations include performance validation primarily on simulated and small-scale experimental data – future work will focus on large-scale real-world deployments across diverse imaging conditions. The

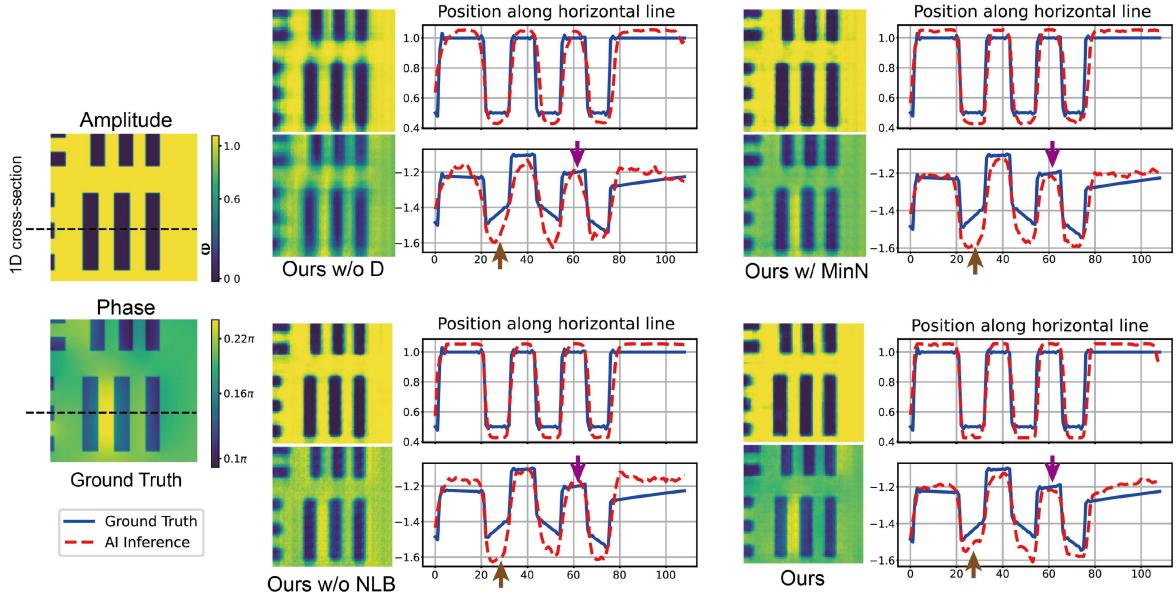

Fig. 9. Ablation study performed on simulated data showing amplitude and phase reconstructions with 2D images and 1D cross-sectional profiles, comparing ground truth (blue solid lines) with AI inference results (red dashed lines).

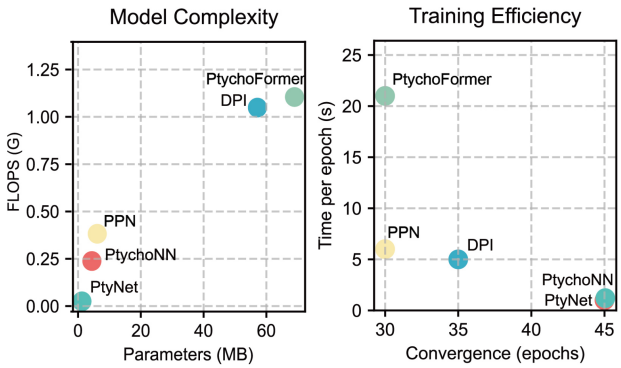

Fig. 10. Comparison of different models in terms of their complexity (parameters and FLOPS) (left) and training efficiency (convergence epochs and time per epoch) (right).

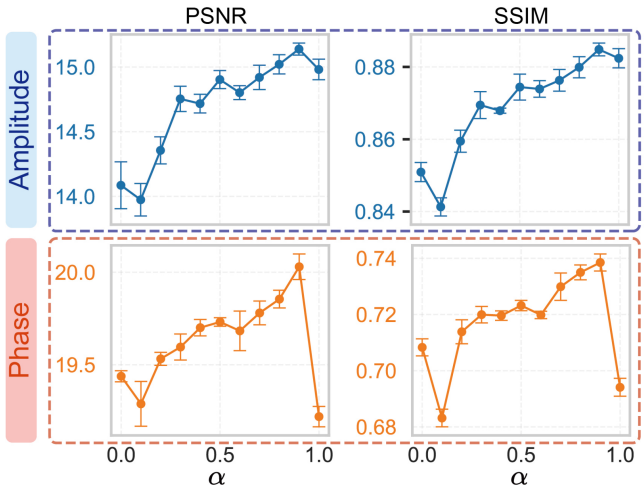

Fig. 11. Performance analysis of weighted loss function $L_{\text{combined}}(\alpha) = \alpha L_{\text{MSE}} + (1-\alpha)L_{\text{NSSIM}}$ on simulated data. PSNR, SSIM are plotted against α for amplitude and phase retrievals.

framework's physics-inspired design principles show promising extensibility to other frequency-domain inverse problems like cryo-EM and astronomical imaging.

ACKNOWLEDGMENT

The authors would like to thank the University of Sydney's Digital Sciences Initiative for financial support of this project through grant DSI Ignite grant scheme.

REFERENCES

[1] R. Hegerl and W. Hoppe, "Dynamische theorie der kristallstrukturanalyse durch elektronenbeugung im inhomogenen primärstrahlwellenfeld," *Berichte der Bunsengesellschaft für physikalische Chemie*, vol. 74, no. 11, pp. 1148–1154, Nov. 1970.

[2] Y. Jiang et al., "Electron ptychography of 2D materials to deep sub-ångström resolution," *Nature*, vol. 559, no. 7714, pp. 343–349, 2018.

[3] W. Yang, H. Sha, J. Cui, L. Mao, and R. Yu, "Local-orbital ptychography for ultrahigh-resolution imaging," *Nature Nanotechnol.*, vol. 19, pp. 612–617, 2024.

[4] E. B. L. Pedersen et al., "Improving organic tandem solar cells based on water-processed nanoparticles by quantitative 3D nanoimaging," *Nanoscale*, vol. 7, no. 32, pp. 13765–13774, 2015.

[5] R. Kasprowicz, R. Suman, and P. O'Toole, "Characterising live cell behaviour: Traditional label-free and quantitative phase imaging approaches," *Int. J. Biochem. Cell Biol.*, vol. 84, pp. 89–95, 2017.

[6] M. Holler et al., "High-resolution non-destructive three-dimensional imaging of integrated circuits," *Nature*, vol. 543, no. 7645, pp. 402–406, 2017.

[7] A. V. Babu et al., "Deep learning at the edge enables real-time streaming ptychographic imaging," *Nature Commun.*, vol. 14, no. 1, 2023, Art. no. 7059.

[8] T. B. Edo et al., "Sampling in X-ray ptychography," *Phys. Rev. A*, vol. 87, no. 5, May 2013, Art. no. 0 53850.

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

[10] D. J. Chang et al., "Deep-learning electron diffractive imaging," *Phys. Rev. Lett.*, vol. 130, no. 1, Jan. 2023, Art. no. 0 16101.

[11] X. Pan et al., "An efficient ptychography reconstruction strategy through fine-tuning of large pre-trained deep learning model," *Iscience*, vol. 26, no. 12, 2023, Art. no. 108420.

[12] M. Deng, S. Li, A. Goy, I. Kang, and G. Barbastathis, "Learning to synthesize: Robust phase retrieval at low photon counts," *Light: Sci. Appl.*, vol. 9, no. 1, 2020, Art. no. 36.
[13] R. Fan et al., "Phase retrieval based on deep learning with bandpass filtering in holographic data storage," *Opt. Exp.*, vol. 32, no. 3, pp. 4498–4510, 2024.
[14] P. M. Pelz, M. Guizar-Sicairos, P. Thibault, I. Johnson, M. Holler, and A. Menzel, "On-the-fly scans for X-ray ptychography," *Appl. Phys. Lett.*, vol. 105, no. 25, 2014, Art. no. 251101.
[15] J. Seifert, Y. Shao, and A. P. Mosk, "Noise-robust latent vector reconstruction in ptychography using deep generative models," *Opt. Exp.*, vol. 32, no. 1, pp. 1020–1033, 2024.
[16] M. Du, X. Huang, and C. Jacobsen, "Using a modified double deep image prior for crosstalk mitigation in multislice ptychography," *J. Synchrotron Radiat.*, vol. 28, no. 4, pp. 1137–1145, 2021.
[17] W. Gan, Q. Zhai, M. T. McCann, C. G. Cardona, U. S. Kamilov, and B. Wohlberg, "PtychoDV: Vision transformer-based deep unrolling network for ptychographic image reconstruction," *IEEE Open J. Signal Process.*, vol. 5, pp. 539–547, 2024.
[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[20] A. Dosovitskiy et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[21] Y. Li et al., "MVitV2: Improved multiscale vision transformers for classification and detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4804–4814.
[22] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.
[23] J. Wu, W. Ji, H. Fu, M. Xu, Y. Jin, and Y. Xu, "MedSegDiff-V2: Diffusion-based medical image segmentation with transformer," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, 2024, pp. 6030–6038.
[24] T. Nguyen, J. Jewik, H. Bansal, P. Sharma, and A. Grover, "ClimateLearn: Benchmarking machine learning for weather and climate modeling," in *Proc. 37th Int. Conf. Neural Inf. Process. Syst.*, 2024, pp. 75009–75025.
[25] J. Chang and J. C. Ye, "Bidirectional generation of structure and properties through a single molecular foundation model," *Nature Commun.*, vol. 15, no. 1, 2024, Art. no. 2323.
[26] R. Nakahata, S. Zaman, M. Zhang, F. Lu, and K. Chiu, "PtychoFormer: A transformer-based model for ptychographic phase retrieval," 2024, *arXiv:2410.17377*.
[27] G. B. Parrent and B. J. Thompson, "On the fraunhofer (far field) diffraction patterns of opaque and transparent objects with coherent background," *Optica Acta: Int. J. Opt.*, vol. 11, no. 3, pp. 183–193, Jul. 1964.
[28] J. Z. Buchwald and C.-P. Yeang, "Kirchhoff's theory for optical diffraction, its predecessor and subsequent development: The resilience of an inconsistent theory," *Arch. Hist. Exact Sci.*, vol. 70, no. 5, pp. 463–511, Sep. 2016.
[29] A. M. Maiden and J. M. Rodenburg, "An improved ptychographical phase retrieval algorithm for diffractive imaging," *Ultramicroscopy*, vol. 109, no. 10, pp. 1256–1262, 2009.
[30] T. Zhou, M. Cherukara, and C. Phatak, "Differential programming enabled functional imaging with Lorentz transmission electron microscopy," *NPJ Comput. Mater.*, vol. 7, no. 1, 2021, Art. no. 141.
[31] A. Bhartiya et al., "X-ray ptychography imaging of human chromosomes after low-dose irradiation," *Chromosome Res.*, vol. 29, no. 1, pp. 107–126, Mar. 2021.
[32] J. N. Weker, X. Huang, and M. F. Toney, "In situ X-ray-based imaging of nano materials," *Curr. Opin. Chem. Eng.*, vol. 12, pp. 14–21, 2016.
[33] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.
[34] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.
[35] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 12077–12090, 2021.
[36] C. Stringer and M. Pachitariu, "Cellpose3: One-click image restoration for improved cellular segmentation," *Nature Methods*, vol. 22, pp. 592–599, 2025.
[37] J. Chen et al., "PIXART-Sigma : Weak-to-strong training of diffusion transformer for 4 k text-to-image generation," in *Proc. Eur. Conf. Comput. Vis.*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham, Switzerland: Springer, 2025, vol. 15090, pp. 74–91.
[38] R. Azad et al., "Medical image segmentation review: The success of u-net," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10076–10095, Dec. 2024.
[39] T. Zhao et al., "A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities," *Nature Methods*, vol. 22, pp. 166–176, 2025.
[40] P. J. Huber, "Robust Estimation of a Location Parameter," in *Breakthroughs in Statistics*, S. Kotz and N. L. Johnson, Eds. New York, NY, USA: Springer New York, 1992, pp. 492–518.
[41] K. Pearson, "VII. Note on regression and inheritance in the case of two parents," *Proc. Roy. Soc. London*, vol. 58, no. 347–352, pp. 240–242, Dec. 1895.
[42] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Jan. 2016.
[43] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.