Learning Physics-based Forward Model Corrections for Unrolled Networks in Diffuser-based Imaging

Yoko Sogabe[1], Shiori Sugimoto[1], Shoichiro Saito[1], and Masaki Kitahara[1]

NTT, Inc., Japan
{yoko.sogabe,shiori.sugimoto,shoichiro.saito,masaki.kitahara}@ntt.com

Abstract. Diffuser-based imaging systems replace conventional lenses with optical diffusers to enable compact lensless cameras, but their performance critically depends on computational reconstruction from compressed measurements. Deep unrolled networks have emerged as a leading framework by incorporating physical forward models into iterative optimization. However, existing methods typically rely on simplified assumptions, most notably shift-invariant point spread functions (PSFs), which fail to capture spatially-varying aberrations and geometric distortions in real hardware. In this paper, we propose learning physics-based forward model corrections for diffuser-based imaging. We construct a physically interpretable forward model with learnable correction terms based on Eigen-Kernel decomposition to handle spatially-varying aberrations. This design preserves the structure of the physical model while improving its fidelity to real measurements. The corrected forward model is integrated into an unrolled reconstruction framework (EK-DUN), leading to consistent improvements in reconstruction accuracy. Experimental results on real-world data demonstrate that our method achieves state-of-the-art performance compared to existing unrolled baselines.

Keywords: Computational imaging · Lensless imaging · Deep Unrolled Networks

1 Introduction

Diffuser-based lensless imaging systems [2, 18] have emerged as a compelling alternative to conventional lens-based cameras, offering the potential for extremely compact, lightweight, and low-cost imaging hardware. By replacing a bulky lens with a thin, scattering optical element (diffuser) placed in front of a sensor, these systems encode the scene information into pseudo-random coded intensity patterns. Recovering the latent image from these captured measurements is, however, a highly ill-posed inverse problem, requiring robust reconstruction algorithms. The unique encoding properties of diffusers have enabled novel applications such as 3D imaging [3], 3D fluorescence microscopy [19], fast video reconstruction [4], and privacy-preserving imaging [16, 22, 31]. However, recovering

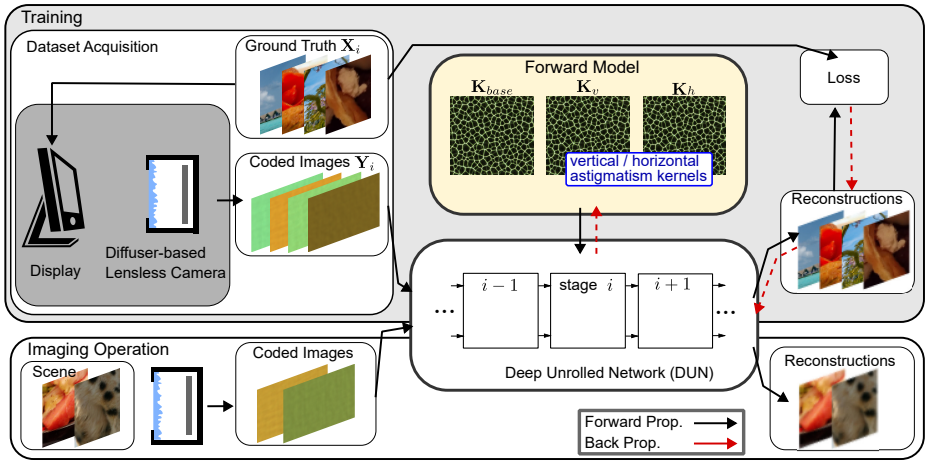

Fig. 1: Overview of the proposed diffuser-based lensless imaging system. First, arbitrary images are displayed on a monitor and captured by the diffuser-based lensless camera to collect a large-scale training dataset $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^{N_D}$ consisting of ground-truth and coded image pairs. Using this dataset $\mathcal{D}$, we train a DUN incorporating the proposed physics-based forward model correction. Specifically, alongside the DUN parameters, the base PSF and the horizontal/vertical eigen-kernels in the forward model are jointly learned to acquire a more physically faithful mapping. During the actual imaging operation, the acquired coded image is fed into the trained model to obtain the reconstructed image.

high-fidelity 2D RGB images from heavily multiplexed measurements remains the foundation of these systems and is the primary focus of this paper.

Deep learning-based approaches have significantly improved reconstruction quality in diffuser-based imaging. In particular, Deep Unrolled Networks (DUNs) have established themselves as a state-of-the-art framework [17, 21, 23, 32, 33]. Unlike purely implicit neural networks, unrolled networks interpret the layers of a neural network as steps in an iterative optimization algorithm, explicitly incorporating a known physical forward model of the system. This improves interpretability and generalization over end-to-end methods.

A critical assumption in most existing unrolled architectures for diffuser-based imaging is the accuracy of the forward model. To ensure computational tractability, prior works typically model the system using a shift-invariant point spread function (PSF), treating the sensing process as a simple 2D convolution. While computationally efficient, this assumption is often violated in real-world hardware setups. Practical optical systems suffer from spatially varying aberrations, geometric distortions, and intensity variations (e.g., vignetting), meaning that the effective PSF differs significantly across the field of view. When the assumed physical model deviates from the actual optical process—a phenomenon we refer to as *forward model mismatch*—the reconstruction performance of un-

rolled networks is severely limited, as the optimization is guided by incorrect physical gradients.

To address this challenge, we propose a novel deep unrolled network, termed Eigen-Kernel Deep Unrolled Network (EK-DUN), which explicitly corrects these physics-based forward model mismatches. Integrating a complex, spatially varying forward model into a DUN is highly non-trivial because the network fundamentally requires the observation model to be strictly linear and computationally efficient in both forward and backward (transpose) evaluations. To satisfy these stringent constraints, our architecture incorporates an *Eigen-Kernel Decomposition* approach. We structurally approximate the uncalibrated, spatially varying PSF as a linear superposition of a base kernel and directional eigen-kernels (representing principal variations like vertical and horizontal astigmatism). By refining the physical model analytically within the unrolling steps, our method effectively bridges the gap between the simplified convolution model and the complex reality of the hardware, maximizing physical fidelity without sacrificing the mathematical tractability required for unrolled optimizations.

Our main contributions are summarized as follows:

1. We conduct a systematic analysis of the forward model mismatch in a real-world diffuser-based camera setup, verifying how the conventional shift-invariant assumption breaks down due to distinct geometric distortions and spatially-varying aberrations.
2. We propose a mathematically tractable and physically interpretable forward model based on Eigen-Kernel decomposition. This formulation enables the unrolled network to learn spatially varying corrections end-to-end while strictly preserving the linear convolution structures and the efficiency of the transpose computations.
3. To balance robust reconstruction and computational efficiency, we introduce an iteration-aware late-activation strategy. By applying the refined Eigen-Kernel model only in the final stages of the unrolled optimization, we effectively resolve high-frequency details, achieving state-of-the-art 2D RGB imaging performance on real-world datasets with minimal computational overhead.

2 Related Work

Diffuser-based imaging is a technology that realizes thin and lightweight camera systems by replacing the lens with other optical elements, such as coded masks or diffusers, and represents a prominent approach in computational imaging. Methods employing optical diffusers [2, 18] or amplitude/phase masks [6, 8] for encoding have been extensively proposed. In this technology, the reconstruction algorithm that recovers the original scene from the heavily multiplexed sensor measurements fundamentally dictates the final image quality. Early studies approached this inverse problem through convex optimization [9] with sparsity constraints, such as Total Variation (TV) minimization [26], based on the compressive sensing framework [11, 12].

Deep Unrolled Networks systematically unfold iterative optimization algorithms (e.g., ADMM [9], ISTA [7]) into deep network layers, effectively integrating physics-based forward models with learned priors. This approach has demonstrated significant success in various inverse problems, including compressive sensing [14, 27, 29, 34], and has become the leading methodology in diffuser-based imaging [17, 21, 23, 32, 33]. However, their performance critically depends on the accuracy of the embedded forward model; mismatches between this mathematical model and real physical hardware inevitably induce artifacts and structural information loss.

Physically faithful forward models are critical for high-quality reconstruction in diffuser-based imaging, yet most methods have largely relied on the idealized assumption that the Point Spread Function (PSF) of the optical system is perfectly constant across the entire field of view (shift-invariant). This allows formulating the forward model purely as a computationally efficient 2D spatial convolution operation with a single, globally constant PSF [21, 23, 32]. Although learning-based methods to estimate the PSF from real data have been proposed [24], they still fundamentally assume a shift-invariant forward model. In practical optical systems, however, the shape and intensity of the PSF vary significantly between on-axis and off-axis regions, causing the shift-invariance assumption to break down and inducing severe model mismatch. To address this mismatch, Zeng et al. [33] proposed compensating for the error a posteriori using a CNN (Compensation Branch) placed in parallel with an ADMM unrolled network that employs an inaccurate shift-invariant model. Furthermore, the Unrolled Primal-Dual Network (UPDN) [17] addresses the breakdown of shift-invariance by introducing 15 learnable PSFs that are maintained as parallel feature maps within the network's intermediate feature space. These representations are not physically interpretable. Similarly, PhoCoLens [10] attempts to address shift-variance by stitching regions processed with different PSFs, though it relies on computationally intensive diffusion models for final quality. To the best of our knowledge, prior unrolled methods for diffuser-based imaging do not embed an explicit and physically interpretable correction for off-axis geometric distortions and continuous shift-variant asymmetric PSF patterns directly into the forward model, while preserving the linear and computationally tractable structure required for unrolled optimization.

3 Preliminaries

3.1 Formulation of Diffuser-based Imaging

Diffuser-based imaging replaces a conventional lens with an optical diffuser placed a few millimeters in front of a standard RGB image sensor. This diffuser can be regarded as a pseudo-random phase mask that modulates the phase of the incoming light. The imaging process in this system, i.e., the raw measurement (coded image) acquired by the sensor for a given scene (target object), can be expressed as a linear observation model:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^N$ is a vector representing the scene ($N = H \times W \times C$, where $C = 3$ for RGB), $\boldsymbol{y} \in \mathbb{R}^M$ is a vector representing the coded image acquired by the sensor ($M = H' \times W' \times C$), and $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ is the measurement matrix of the system. Throughout the paper, we use the term *forward model* to refer to this linear measurement operator $\boldsymbol{A}$ and its efficient implementation (e.g., convolution under shift-invariance).

The Point Spread Function (PSF) of diffuser-based imaging forms a large, high-contrast caustic pattern that spreads across the entire field of view (see Fig. 1). With appropriate hardware design (e.g., optimizing the diffusion angle of the diffuser and the distance between the sensor and the diffuser), the PSF of diffuser-based imaging can be approximately modeled as spatially shift-invariant [21]. That is, when a point light source in the scene shifts laterally, the PSF pattern on the sensor translates in the opposite direction without changing its shape. By introducing this shift-invariance assumption, the multiplication by the measurement matrix $\boldsymbol{A}$ can be efficiently described as a 2D convolution operation using a 3rd-order tensor representation:

$$\mathbf{Y} = \mathbf{K} * \mathbf{X} \tag{2}$$

where $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ is the tensor (or image) representation of the scene, $\mathbf{Y} \in \mathbb{R}^{H' \times W' \times C}$ is the tensor representation of the coded image, $\mathbf{K} \in \mathbb{R}^{H_k \times W_k \times C}$ is the PSF (kernel) of the system, and $*$ denotes the 2D discrete convolution operation in the spatial domain. In this paper, we denote vectors as bold lowercase letters (e.g., $\boldsymbol{x}, \boldsymbol{y}$) and their 2D/3D tensor representations equivalently as bold uppercase letters (e.g., $\mathbf{X}, \mathbf{Y}$) depending on the context.

Since the measurement matrix $\boldsymbol{A}$ is extremely large ($M \times N$), direct matrix-vector multiplication is impractical in memory and compute. Under shift-invariance, $\boldsymbol{A}$ can be implemented as convolution (Eq. (2)), enabling efficient forward computation and an exact adjoint. For this reason, the shift-invariant convolutional model is widely adopted in prior diffuser-based reconstruction methods [4, 21, 23, 32, 33].

3.2 Inverse Problem & Reconstruction with Unrolled Networks

In diffuser-based imaging, the raw sensor measurements are heavily multiplexed. Therefore, to recover the spatial structure of the original scene $\boldsymbol{x}$ from the coded image $\boldsymbol{y}$ for visual inspection or downstream vision tasks, a computational reconstruction process (solving an inverse problem) is essential. This inverse problem is generally formulated as an optimization problem with regularization:

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \tau R(\boldsymbol{x}) \right\} \tag{3}$$

where the first term on the right-hand side is the data fidelity term, ensuring that the solution matches the observed data. The second term, $R(\boldsymbol{x})$, is a prior/regularization term based on prior knowledge, and τ is a parameter that balances the two terms.

Traditionally, such problems are solved via convex optimization using priors like Total Variation (TV) [26] and algorithms like the Alternating Direction Method of Multipliers (ADMM) [9]. In contrast, DUNs unfold the iterative processes of classical optimization algorithms (such as ADMM or ISTA [7]) into layers of a neural network, providing a powerful method that combines iterative mathematical optimization with learned image priors while explicitly incorporating the physical forward model.

Taking an ADMM-based DUN [27] as an example, by introducing variable splitting, the solution is obtained by iteratively computing the following update equations:

$$\begin{aligned}
\boldsymbol{g}^{(i+1)} &= \underset{\boldsymbol{g}}{\operatorname{argmin}} \left\{ \frac{1}{2}\|\boldsymbol{g} - (\boldsymbol{f}^{(i)} - \boldsymbol{d}^{(i)})\|_2^2 + \frac{\tau}{\eta^{(i)}} R(\boldsymbol{g}) \right\} \\
&= \mathbb{P}^{(i)}(\boldsymbol{f}^{(i)} - \boldsymbol{d}^{(i)}) \qquad (4) \\
\boldsymbol{f}^{(i+1)} &= \underset{\boldsymbol{f}}{\operatorname{argmin}} \left\{ \frac{1}{2}\|\boldsymbol{A}\boldsymbol{f} - \boldsymbol{y}\|_2^2 + \frac{\eta^{(i)}}{2}\|\boldsymbol{f} - (\boldsymbol{g}^{(i+1)} + \boldsymbol{d}^{(i)})\|_2^2 \right\} \\
&\approx \left[(1 - \epsilon^{(i)}\eta^{(i)})\boldsymbol{I} - \epsilon^{(i)}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A}\right] \boldsymbol{f}^{(i)} + \epsilon^{(i)}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{y} + \epsilon^{(i)}\eta^{(i)}(\boldsymbol{g}^{(i+1)} + \boldsymbol{d}^{(i)}) \qquad (5) \\
\boldsymbol{d}^{(i+1)} &= \boldsymbol{d}^{(i)} + \kappa^{(i)}(\boldsymbol{g}^{(i+1)} - \boldsymbol{f}^{(i+1)}) \qquad (6)
\end{aligned}$$

where i is the iteration number, $\boldsymbol{f} \in \mathbb{R}^N$ is the iterative estimate (initialized with $\boldsymbol{f}^{(0)} = \boldsymbol{A}^{\mathsf{T}}\boldsymbol{y}$), $\boldsymbol{g} \in \mathbb{R}^N$ is an auxiliary variable, and $\boldsymbol{d} \in \mathbb{R}^N$ is a scaled dual variable initialized to $\mathbf{0}$. The final reconstructed image is given by the output of the last iteration, $\hat{\boldsymbol{x}} = \boldsymbol{f}^{(I)}$, where I is the total number of iterations. $\epsilon^{(i)}, \eta^{(i)}, \kappa^{(i)} \in \mathbb{R}$ are the step size of the gradient method, the penalty parameter, and the update rate of the dual variable, respectively, which are defined as learnable parameters specific to each iteration. $\mathbb{P}^{(i)}$ in Eq. (4) represents a proximal mapping operator related to the regularization. The Plug-and-Play (PnP) framework [28] has elegantly demonstrated that this proximal operator does not need to be derived analytically from an explicit convex function; rather, it can be replaced by an implicit, data-driven image filter or denoiser. Building upon this insight, DUNs replace $\mathbb{P}^{(i)}$ with a learnable deep neural network (DNN). In this paper, we refer to this network $\mathbb{P}^{(i)}$ as the prior network. That is, in DUNs, the optimization hyperparameters $\epsilon^{(i)}, \eta^{(i)}, \kappa^{(i)}$ and the weights of the prior network $\mathbb{P}^{(i)}$ are defined independently for each iteration and optimized end-to-end from training data. As shown in Fig. 1, capturing displayed images enables scalable data collection ($\mathcal{D} = \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^{N_D}$), strongly supporting this data-driven optimization.

Requirements for Forward Models in DUNs. To integrate a forward model (Eq. (1)) into the explicit physics-based update of DUNs (e.g., Eq. (5)), it should ideally satisfy three key requirements:

- **Physical Fidelity:** The model must accurately describe the actual light transport process. Model mismatches directly degrade reconstruction quality.
- **Linearity:** The forward model in DUNs fundamentally assumes a linear observation model (Eq. (1)). Introducing non-linear structures (e.g., CNNs with

activations) directly into the forward computation breaks this mathematical foundation.
- **Tractable Adjoint:** The iterative computation of $\boldsymbol{Ax}$ and its exact adjoint $\boldsymbol{A}^\top \boldsymbol{y}$, which appears in the DUN update in Eq. (5), must be extremely lightweight and analytically derivable. This makes implicit models, such as deep neural networks, generally unsuitable.

The standard shift-invariant convolutional model (Eq. (2)) has been widely adopted [17, 21, 23, 32, 33] because it perfectly satisfies the mathematical requirements (linearity and computationally lightweight exact adjoint via cross-correlation), despite lacking sufficient physical fidelity in real-world setups.

4 Proposed Method

To address severe model mismatch in real hardware (Sec. 4.1), we propose an Eigen-Kernel decomposition approach. We formulate interpretable, shift-variant PSF corrections preserving the tractability required for DUNs (Sec. 4.2) and integrate them into the iterative optimization (Sec. 4.3).

4.1 Analysis of Shift-Invariance in Real-World Setup

We conduct a systematic analysis of the PSF in our real-world diffuser-based imaging setup (*cf*. Sec. 5.1) to verify the breakdown of the shift-invariance assumption. Figure 2 shows the captured PSF patterns at different positions of a point light source in the scene. By taking the difference between the PSF at the optical axis center and those at off-axis positions, we observe distinct errors. Specifically, we identify two major types of model mismatch:

(i) **Geometric Distortion**: The caustic patterns exhibit subtle geometric distortions (e.g., local warping or stretching) that depend on the incident angle. These distortions prevent a single kernel from perfectly matching the effective PSF at all locations.
(ii) **Global Variations**: As the light source moves away from the optical axis, global intensity variations (gain) and shifts become more pronounced.

Our quantitative analysis further confirms that the Mean Absolute Error (MAE) of the PSF relative to the center PSF increases significantly toward the periphery. These findings indicate that while the shift-invariant assumption is a reasonable first-order approximation, it fails to capture the fine-grained physical variations necessary for high-fidelity reconstruction. For a comprehensive analysis illustrating the full 9×9 grid measurement of shift-variant PSFs, please refer to the supplementary material.

4.2 Proposed Forward Model

To address the limitations of the single shift-invariant kernel assumption, typically caused by spatially varying aberrations such as astigmatism, we propose

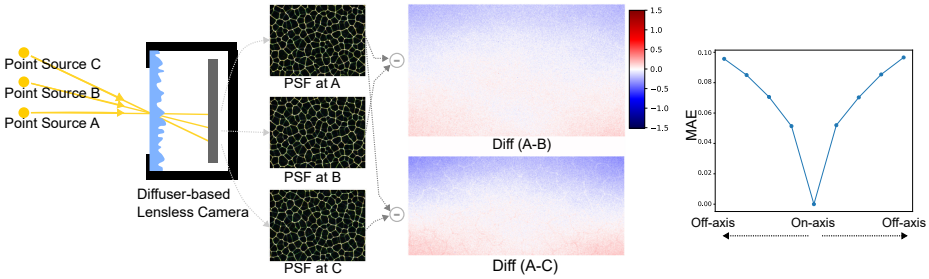

Fig. 2: Breakdown of PSF shift-invariance. We captured PSFs of a vertically shifted point source (A-C). Difference images between the central (A) and off-axis PSFs (B, C) reveal (i) geometric distortions in caustics, visible as residual caustic patterns (particularly in Diff(A-C)), and (ii) global gain variations, which worsen away from the optical axis. The plotted Mean Absolute Error (MAE) relative to the center confirms that shift-invariant modeling errors increase significantly toward the periphery.

a refined forward model based on **Eigen-Kernel Decomposition**. In practical optical systems, the spatially varying PSF $\mathbf{K}(u, v) \in \mathbb{R}^{H_k \times W_k \times C}$ continuously changes depending on the pixel coordinate (u, v) on the sensor. Our approach is grounded in the observation that these physical variations are highly constrained and effectively lie within a low-dimensional subspace. Prior studies have shown that spatially varying lens PSFs (e.g., due to defocus and aberrations) can be accurately represented by a low-dimensional basis obtained via PCA, expressing pixel-wise blur as a linear combination of a small number of eigen-kernels [13,15]. Based on this insight, we approximate the spatially varying PSF by decomposing it into a small number of "eigen-kernels" (principal components), rather than attempting to model pixel-wise unique kernels, which would be computationally prohibitive. Specifically, we hypothesize that the dominant PSF variations can be efficiently represented by a base kernel and two directional difference kernels corresponding to vertical and horizontal astigmatism. Our proposed model, which we term **Eigen-Kernel Deep Unrolled Network (EK-DUN)**, formulates the measurement $\mathbf{Y}$ as a linear combination of convolutions:

$$\mathbf{Y} \approx \underbrace{(\mathbf{1} - \mathbf{W}_h - \mathbf{W}_v) \odot (\mathbf{K}_{base} * \mathbf{X})}_{\text{Mean PSF}} + \underbrace{\mathbf{W}_v \odot (\mathbf{K}_v * \mathbf{X})}_{\text{Vertical Astigmatism}} + \underbrace{\mathbf{W}_h \odot (\mathbf{K}_h * \mathbf{X})}_{\text{Horizontal Astigmatism}} \tag{7}$$

where $\mathbf{K}_{base}$ represents the spatially averaged mean PSF, while $\mathbf{K}_v$ and $\mathbf{K}_h$ are the "eigen-kernels" capturing the principal variations in the vertical and horizontal directions, respectively. The spatial weight maps $\mathbf{W}_v, \mathbf{W}_h \in [0, 1]^{H' \times W'}$ determine the mixing ratio of these kernels at each pixel location (u, v). Instead of learning these weights, we define them analytically to limit the degrees of freedom and prevent the spatial weights from absorbing PSF variations, ensuring that the learned kernels capture meaningful PSF modes. Based on the assumption that aberrations increase symmetrically away from the optical axis, we typically use sinusoidal weights such as $\mathbf{W}_h = \frac{1}{2}\sin^2(0.5\pi u_{norm})$ and

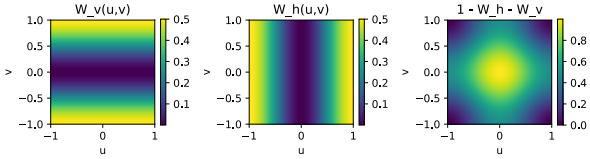

Fig. 3: Visualization of the spatial weight maps $\mathbf{W}_h$, $\mathbf{W}_v$, and $(\mathbf{1} - \mathbf{W}_h - \mathbf{W}_v)$. These maps act as spatial priors to control the mixing of eigen-kernels based on the distance from the optical center along the horizontal and vertical directions.

$\mathbf{W}_v = \frac{1}{2}\sin^2(0.5\pi v_{norm})$, where u_{norm}, v_{norm} are normalized coordinates. Under this design, the further a point is from the optical center along the horizontal or vertical axis, the more strongly $\mathbf{K}_h$ or $\mathbf{K}_v$ compensates for the severe PSF distortion. Crucially, correcting distortions in off-axis regions enhances not only local but global image fidelity. Since light from these regions is multiplexed across the entire sensor, local model mismatches propagate as global artifacts. Figure 3 visualizes these spatial priors and their superposition.

The adjoint operator of Eq. (7) is straightforward to derive because it is a weighted sum of linear convolutions. This preserves the lightweight and analytically derivable adjoint required by the DUN update. Specifically, we first multiply the input by each fixed spatial weight map ($(\mathbf{1} - \mathbf{W}_h - \mathbf{W}_v)$, $\mathbf{W}_v$, and $\mathbf{W}_h$), apply the adjoint convolution with the corresponding kernel, and then sum the three resulting tensors. Both Eq. (7) and its adjoint can be efficiently implemented using FFTs.

This Eigen-Kernel approach allows us to model continuous PSF variations, such as astigmatic stretching, without the blocking artifacts associated with patch-wise approximations. Crucially, this formulation perfectly satisfies the three essential requirements for DUN forward models (Sec. 3): it enhances **physical fidelity** by modeling spatially-varying aberrations, strictly preserves **linearity** through convolutions and element-wise multiplications, and ensures a **tractable adjoint** analytically derivable as a simple combination of FFT-based cross-correlations and element-wise multiplications by the weight maps. This theoretically grounded dimensionality reduction achieves high-fidelity modeling with minimal additional kernels, avoiding the overfitting and computational burden associated with pixel-wise or higher-order approximations.

4.3 Integration into Unrolled Network

We integrate this Eigen-Kernel forward model into the ADMM-based Deep Unrolled Network (Eqs. (4) to (6)) as illustrated in Fig. 4. While accurate, the multi-kernel model is more computationally expensive than the standard single-kernel model. To balance performance and efficiency, we adopt a hybrid strategy, hypothesizing that the refined physical model is most critical in the final stages where fine details are resolved. Therefore, we apply the proposed Eigen-Kernel model (Eq. (7)) and its adjoint only in the **last 2 iterations** of the unrolled

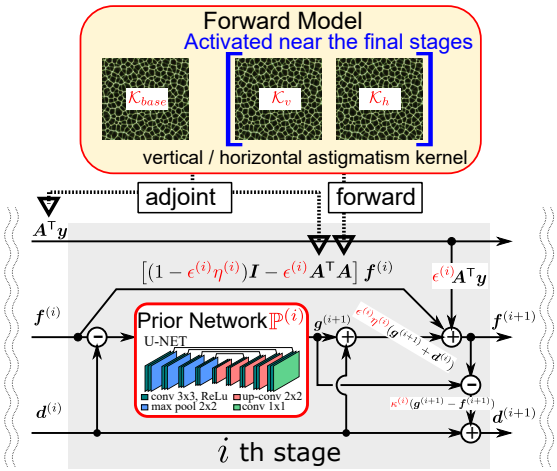

Fig. 4: Overview of the i-th iteration stage in our proposed Deep Unrolled Network. In the early stages, only the base kernel $\mathbf{K}_{base}$ is used for the physical model update to maintain efficiency. In the final stages, the refined Eigen-Kernel model (incorporating $\mathbf{K}_v$ and $\mathbf{K}_h$) is activated to correct for residual asymmetric aberrations. Components highlighted in red indicate learnable parameters.

network. All preceding iterations use the standard shift-invariant model with the shared base kernel $\mathbf{K}_{base}$. In this framework, the kernels $\mathbf{K}_{base}, \mathbf{K}_v, \mathbf{K}_h$ are optimized end-to-end along with the unrolled network parameters, allowing the system to learn the optimal physical decomposition from data.

5 Experiments

5.1 Experimental Setup

We evaluate the proposed method using a custom-built diffuser-based lensless camera prototype assembled from off-the-shelf optical components, as shown in Fig. 5. We utilize a 1.6-megapixel CMOS sensor equipped with a 1° Engineered Diffuser. Detailed specifications and the construction guide for this hardware setup are provided in the supplementary material. To obtain a large-scale dataset for training and evaluation, we fix the camera position to face a monitor. The field of view (FoV) of the camera relative to the monitor is determined to be 442×614 pixels by displaying white points at various positions.

We curate a dataset using high-quality images from BSDS500 [5], DIV2K [1], and Flickr2K [20]. The images are strictly divided by image ID into training and test sets to prevent data leakage. During data collection, images are randomly cropped to the predetermined FoV and displayed on the monitor, while the camera captures the corresponding coded images at a resolution of 1440×1080. In total, we collect 32,000 image pairs for training and 2,000 for testing. The

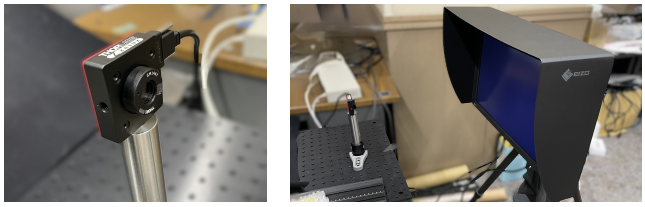

Fig. 5: Experimental setup. Left: Close-up view of the lensless camera prototype (zoom shot). Right: Overview of the data collection setup with the monitor (overview shot). We construct a lensless camera using a generic image sensor and a diffuser, capturing coded images displayed on a monitor placed at a distance of approximately 50 cm.

system point spread function (PSF) is experimentally measured by capturing the response to a single white point displayed at the center of the monitor.

We compare our method against two representative baselines: the standard Le-ADMM-U [21] and the state-of-the-art UPDN (15 mixed channels) [17]. Although a more recent approach, PhoCoLens [10], employs conditional diffusion models to enhance subjective perceptual quality, it entails prohibitive computational costs and, as noted by its authors, yields lower PSNR than UPDN. For this evaluation, UPDN is therefore treated as the strongest baseline for fidelity-oriented metrics such as PSNR.

5.2 Implementation Details

The proposed EK-DUN is implemented with 10 iterations. For the prior network $\mathbb{P}^{(i)}$ in each iteration, we employ a U-Net [25]. Specifically, we configure a 3-level U-Net where each level consists of two standard convolution blocks (Conv-ReLU), max pooling for downsampling, and transposed convolutions for upsampling. We use 32 base feature channels that double at each subsequent level. All models are trained to minimize the Mean Squared Error (MSE) loss under identical conditions for 200 epochs using an initial learning rate of 1×10^{-4} and an exponential learning rate scheduler with a decay factor of 0.95 per epoch. We follow the default settings reported in prior work for each baseline (e.g., Le-ADMM-U with 5 iterations and UPDN with 10 iterations).

The experimentally measured PSF is provided to all models. In the Le-ADMM-U baseline, the PSF is kept fixed as a constant physical parameter. In contrast, for both UPDN and our proposed method, the PSF is initialized with the measured data and subsequently optimized jointly with the network parameters. All models were trained and evaluated on a single NVIDIA A100 (80GB) GPU with PyTorch 2.1.0 and CUDA 11.8.

5.3 Qualitative Results

Figure 6 provides a qualitative comparison of the reconstruction results on the test set. The baseline Le-ADMM-U, based on a fixed shift-invariant PSF, suffers from noticeable blurring and artifacts, particularly in regions far from the

Fig. 6: Qualitative comparison of reconstruction results on real-world data. We compare EK-DUN with Le-ADMM-U, a baseline unrolled network with a shift-invariant PSF, and UPDN, a state-of-the-art unrolled network with multiple learned PSFs. EK-DUN achieves high-fidelity reconstruction comparable to UPDN, preserving fine details and reducing artifacts in challenging regions.

Table 1: Quantitative comparison of reconstruction performance on the test set. We report the average PSNR, SSIM, and LPIPS. The best results are highlighted in **bold**, and the second best are underlined.

Method	PSFs	# Iter.	# Params	Runtime [s] ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Le-ADMM-U [21]	1 RGB (fixed)	5	38.0M	**0.132**	27.65	0.810	0.2994
UPDN [17]	15 mixed (trainable)	10	**11.9M**	0.257	<u>30.65</u>	<u>0.837</u>	<u>0.2764</u>
Ours (EK-DUN)	3 RGB (trainable)	10	82.7M	<u>0.187</u>	**31.09**	**0.843**	**0.2650**

optical axis where PSF variations are significant. UPDN significantly improves sharpness by learning multiple PSFs to approximate shift-variance. Our proposed method achieves comparable visual quality to this state-of-the-art baseline, effectively resolving fine details and textures. This result demonstrates that our physically grounded Eigen-Kernel approach successfully captures the shift-variant optics necessary for high-quality imaging, matching the performance of complex learned-filter baselines while maintaining physical interpretability.

5.4 Quantitative Performance

We evaluate the reconstruction quality using three standard metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [30], and Learned Perceptual Image Patch Similarity (LPIPS) [35]. Table 1 summarizes the quantitative performance. Our EK-DUN significantly outperforms Le-ADMM-U (+3.44 dB) and surpasses the state-of-the-art UPDN by +0.44 dB in PSNR. Notably, EK-DUN achieves this quality with a faster inference runtime (0.187 s vs 0.257 s for UPDN). Although EK-DUN has more trainable parameters than UPDN, parameter count is not the dominant inference cost in this setting. The main cost comes from FFT-based convolutions with large PSFs and their intermediate feature maps; UPDN employs 5× more kernels per layer to handle shift-variance, whereas EK-DUN applies only a small number of additional kernels in the final stages. Consistent with this, EK-DUN also reduces peak inference GPU memory compared with UPDN (2.03 GiB vs 2.86 GiB). Thus, EK-DUN improves PSNR over UPDN while also reducing inference runtime and peak memory, despite having a larger number of trainable parameters. It also preserves a strictly linear forward model with physically interpretable corrections.

5.5 Ablation Study

To verify the effectiveness of our proposed hybrid Eigen-Kernel strategy, we compare three EK-DUN variants with different forward-model configurations: (1) a $\mathbf{K}_{base}$-only variant, which corresponds to a shift-invariant unrolled network with the same 10-iteration architecture, (2) a full Eigen-Kernel variant that applies the Eigen-Kernel model ($\mathbf{K}_{base}, \mathbf{K}_v, \mathbf{K}_h$) in all 10 iterations, and (3) our proposed hybrid variant that activates the Eigen-Kernel model only in the last 2 iterations.

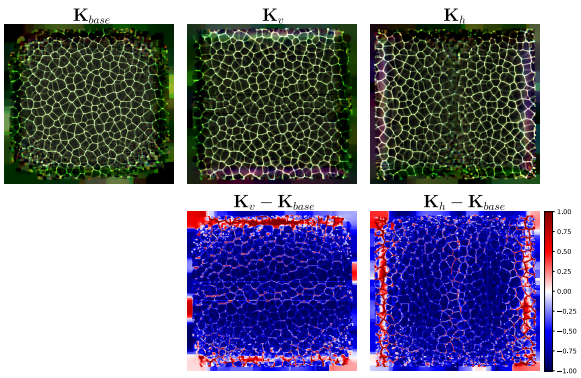

Fig. 7: Visualization of the learned base kernel $\mathbf{K}_{base}$, vertical eigen-kernel $\mathbf{K}_v$, horizontal eigen-kernel $\mathbf{K}_h$, and their respective differences from the base. The difference images reveal subtle displacements in the caustic spot positions, indicating that the kernels have learned to adaptively compensate for spatially-varying vertical and horizontal distortions.

Table 2: Ablation study of different kernel configurations. We report the average PSNR on the test set and the inference runtime in seconds per image.

Configuration	PSNR ↑	Runtime [s] ↓
EK-DUN w/ $\mathbf{K}_{base}$ only	29.54	**0.181**
EK-DUN w/ Eigen-Kernel (all iterations)	**31.12**	0.246
EK-DUN w/ Eigen-Kernel (last 2 iterations)	<u>31.09</u>	<u>0.187</u>

Table 2 summarizes the results. The improvement from the $\mathbf{K}_{base}$-only variant to the full Eigen-Kernel variant reflects the benefit of replacing the shift-invariant forward model with the proposed Eigen-Kernel forward model. While applying the full Eigen-Kernel model to all iterations yields the highest PSNR, it increases the inference runtime by approx. 36%. The small PSNR gap between the full and hybrid variants supports the hypothesis introduced in Sec. 4.3: the refined forward model is most beneficial in the final reconstruction stages, where remaining errors are dominated by fine spatial details. The hybrid variant therefore retains almost all of the PSNR gain of the full Eigen-Kernel variant while remaining close to the $\mathbf{K}_{base}$-only runtime.

5.6 Visualization of Learned Eigen-Kernels

To provide physical insight into the learned forward model corrections, we visualize the base kernel $\mathbf{K}_{base}$ and the directional eigen-kernels $\mathbf{K}_v, \mathbf{K}_h$ in Fig. 7. By observing the difference images ($\mathbf{K}_v - \mathbf{K}_{base}$ and $\mathbf{K}_h - \mathbf{K}_{base}$), we find that the network effectively learns to shift and warp the caustic patterns. Specifically, the difference maps exhibit subtle sub-pixel displacements of the spot structures

along the vertical and horizontal axes, respectively. These results confirm that the unrolled network succeeds in acquiring spatially-varying physical parameters that directly compensate for the geometric and astigmatic distortions observed in our hardware setup (Sec. 4.1).

6 Conclusion

In this paper, we proposed Eigen-Kernel Deep Unrolled Network (EK-DUN), a deep unrolled network for diffuser-based lensless imaging that learns physics-based corrections to the forward model. EK-DUN mitigates shift-invariant model mismatch by decomposing spatially varying PSFs into interpretable eigen-kernels and integrating them into the iterative reconstruction process. Experiments on real data from our prototype show improved reconstruction fidelity over prior unrolled baselines, demonstrating the effectiveness of forward-model mismatch correction in our diffuser-based lensless imaging setup. A remaining limitation is that the current low-dimensional Eigen-Kernel basis may be insufficient for wider FoV diffuser configurations, where larger sensors, larger effective apertures, or wider scattering angles can introduce higher-order off-axis aberrations. Beyond this specific prototype, by learning correction kernels from captured data rather than relying on a detailed hand-crafted aberration model for a specific diffuser, this approach may extend data-driven forward-model corrections to other lensless imaging systems, such as mask-based cameras, and to aberration-limited computational imaging systems including DOE- and metalens-based imaging. More broadly, because the proposed correction preserves a linear forward model with an efficient adjoint operator, its application to non-unrolled reconstruction methods that rely on linear inverse or adjoint computations is also a promising extension for future work.

References

1. Agustsson, E., Timofte, R.: NTIRE 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (July 2017)
2. Antipa, N., Necula, S., Ng, R., Waller, L.: Single-shot diffuser-encoded light field imaging. In: 2016 IEEE International Conference on Computational Photography (ICCP). pp. 1–11. IEEE (2016)
3. Antipa, N., Kuo, G., Heckel, R., Mildenhall, B., Bostan, E., Ng, R., Waller, L.: DiffuserCam: lensless single-exposure 3D imaging. Optica **5**(1), 1–9 (2018)
4. Antipa, N., Oare, P., Bostan, E., Ng, R., Waller, L.: Video from stills: Lensless imaging with rolling shutter. In: 2019 IEEE International Conference on Computational Photography (ICCP). pp. 1–8. IEEE (2019)
5. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE Transactions on pattern analysis and machine intelligence **33**(5), 898–916 (May 2011)
6. Asif, M.S., Ayremlou, A., Sankaranarayanan, A., Veeraraghavan, A., Baraniuk, R.: FlatCam: Thin, bare-sensor cameras using coded aperture and computation. arXiv preprint arXiv:1509.00116 (2015)
7. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM journal on imaging sciences **2**(1), 183–202 (2009)
8. Boominathan, V., Robinson, J.T., Waller, L., Veeraraghavan, A.: Recent advances in lensless imaging. Optica **9**(1), 1–16 (2022)
9. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine learning **3**(1), 1–122 (2011)
10. Cai, X., You, Z., Zhang, H., Gu, J., Liu, W., Xue, T.: PhoCoLens: Photorealistic and consistent reconstruction in lensless imaging. Advances in Neural Information Processing Systems **37**, 12219–12242 (2024)
11. Candès, E.J., Wakin, M.B.: An introduction to compressive sampling. IEEE signal processing magazine **25**(2), 21–30 (2008)
12. Donoho, D.L.: Compressed sensing. IEEE Transactions on information theory **52**(4), 1289–1306 (2006)
13. Gwak, M., Yang, S.: Modeling nonstationary lens blur using eigen blur kernels for restoration. Opt. Express **28**(26), 39501–39523 (Dec 2020)
14. Jacome, R., Bacca, J., Arguello, H.: D^2UF: Deep coded aperture design and unrolling algorithm for compressive spectral image fusion. IEEE Journal of Selected Topics in Signal Processing **17**(2), 502–512 (2023)
15. Jee, M., Blakeslee, J., Sirianni, M., Martel, A., White, R., Ford, H.: Principal component analysis of the time-and position-dependent point-spread function of the advanced camera for surveys. Publications of the Astronomical Society of the Pacific **119**(862), 1403–1419 (2007)
16. Khan, S.S., Yu, X., Mitra, K., Chandraker, M., Pittaluga, F.: OpEnCam: Lensless optical encryption camera. IEEE Transactions on Computational Imaging (2024)
17. Kingshott, O., Antipa, N., Bostan, E., Akşit, K.: Unrolled primal-dual networks for lensless cameras. Optics Express **30**(26), 46324–46335 (2022)
18. Kuo, G., Antipa, N., Ng, R., Waller, L.: DiffuserCam: diffuser-based lensless cameras. In: Computational Optical Sensing and Imaging. pp. CTu3B–2. Optica Publishing Group (2017)

19. Kuo, G., Antipa, N., Ng, R., Waller, L.: 3D fluorescence microscopy with diffusercam. In: Computational Optical Sensing and Imaging. pp. CM3E–3. Optica Publishing Group (2018)
20. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017)
21. Monakhova, K., Yurtsever, J., Kuo, G., Antipa, N., Yanny, K., Waller, L.: Learned reconstructions for practical mask-based lensless imaging. Optics express **27**(20), 28075–28090 (2019)
22. Nguyen Canh, T., Nagahara, H.: Deep compressive sensing for visual privacy protection in FlatCam imaging. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 3978–3986 (2019)
23. Poudel, A., Nakarmi, U.: Deeplir: Attention-based approach for mask-based lensless image reconstruction. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 431–439 (2024)
24. Rego, J.D., Kulkarni, K., Jayasuriya, S.: Robust lensless image reconstruction via PSF estimation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 403–412 (2021)
25. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI. Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer (2015)
26. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: nonlinear phenomena **60**(1-4), 259–268 (1992)
27. Sogabe, Y., Sugimoto, S., Kurozumi, T., Kimata, H.: ADMM-inspired reconstruction network for compressive spectral imaging. In: 2020 IEEE International Conference on Image Processing (ICIP). pp. 2865–2869. IEEE (2020)
28. Venkatakrishnan, S.V., Bouman, C.A., Wohlberg, B.: Plug-and-play priors for model based reconstruction. In: 2013 IEEE global conference on signal and information processing. pp. 945–948. IEEE (2013)
29. Wang, L., Sun, C., Fu, Y., Kim, M.H., Huang, H.: Hyperspectral image reconstruction using a deep spatial-spectral prior. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8032–8041 (2019)
30. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
31. Wang, Z.W., Vineet, V., Pittaluga, F., Sinha, S.N., Cossairt, O., Bing Kang, S.: Privacy-preserving action recognition using coded aperture videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2019)
32. Yang, J., Yin, X., Zhang, M., Yue, H., Cui, X., Yue, H.: Learning image formation and regularization in unrolling AMP for lensless image reconstruction. IEEE Transactions on Computational Imaging **8**, 479–489 (2022)
33. Zeng, T., Lam, E.Y.: Robust reconstruction with deep learning to handle model mismatch in lensless imaging. IEEE Transactions on Computational Imaging **7**, 1080–1092 (2021)
34. Zhang, J., Ghanem, B.: ISTA-Net: Interpretable optimization-inspired deep network for image compressive sensing. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1828–1837 (2018)

35. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)