INTEGRATED FORWARD–INVERSE NETWORKS FOR PHYSICS-GUIDED IMAGE RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Inverse modeling plays a central role across computational optical imaging problems, including microscopy, imaging through scattering media, and lensless cameras, where the forward model often manifests as a severe blur. Discrepancies between the model and the actual imaging process further aggravate the ill-posed nature of the inverse problem. Physics-enabled methods that integrate analytical forward models with data-driven networks have been explored, but most incorporate physics only in a one-sided manner—either operating purely in the measurement space or only after inversion—thereby discarding complementary cues and reducing robustness to calibration errors. Here, we propose the Integrated Forward–Inverse Network (IFIN), a physics-guided deep neural network that interleaves differentiable forward operators with learnable inverse modules at every stage of the hierarchy. This design preserves physical consistency while shaping richer feature representations by jointly leveraging information from both measurement and image domains. A physics-guided kernel adaptation further compensates for inaccurate or unavailable PSF calibration, dynamically refining the kernel for blind deconvolution under system constraints. IFIN is especially effective when measurements are severely blurred by large point-spread functions, where conventional CNN-based inversion is limited by local receptive fields and underutilizes the measurement signal. On challenging lensless imaging benchmarks—including our newly introduced dataset, IFIN achieves state-of-the-art reconstruction quality and improved robustness under noise and model mismatch.

1 INTRODUCTION

Modern optical imaging systems—ranging from compact lensless cameras with coded apertures to advanced microscopes with engineered point-spread functions (PSFs)—are increasingly designed with complex forward models. These systems often operate in regimes where the PSF is intentionally or unavoidably broadened by scattering, diffraction, or designed optical coding, while producing significantly blurred measurements. Such designs unlock diverse imaging capabilities (Sahoo et al., 2017; Satat et al., 2017; Antipa et al., 2017; 2019; Baek et al., 2022) that transcend the limits of conventional optics, yet they also introduce substantial challenges for reconstruction. In particular, the resulting measurements frequently violate the stationarity assumptions underlying standard inverse pipelines (Kuo et al., 2020; Cai et al., 2024), while hardware imperfections and residual modulations further complicate the model. As optical platforms continue to shrink and diversify, the model mismatch between the assumed forward model and the actual device increases, making accurate and robust image restoration a central difficulty.

A wide spectrum of approaches has been explored for image reconstruction under complex forward models. Classical, analytically derived inverse mappings (Wiener, 1964) and model-based optimization (Richardson, 1972; Lucy, 1974; Boyd et al., 2011) built on well-defined priors offer stable and physically valid results, but the methods are often computationally expensive, sensitive to calibration errors, and unreliable under model mismatch. With advances in deep learning, data-driven methods (Anonymous, 2020; Pan et al., 2022) has enabled end-to-end mappings from measurements to images. While such models provide fast inference and can be trained directly on task metrics, they may not explicitly encode the underlying system physics, which can reduce accuracy and robustness under out-of-distribution conditions and occasionally yield hallucinations.

In response, hybrid methods that embed the physical forward model within a learning framework have emerged, improving efficiency and grounding predictions in the physical model while leveraging data-driven components to capture priors that are difficult to specify analytically. Yet in practice, the methods typically integrate physics only in a one-sided manner—for instance, by applying a learned denoiser after physics-based reconstruction—either closed-form or optimization-based—with or without learnable parameters (Monakhova et al., 2019; Khan et al., 2020; Yanny et al., 2022; Kingshott et al., 2022; Poudel & Nakarmi, 2024; Lee et al., 2023a), incorporating the forward model into the loss term (Ulyanov et al., 2018; Monakhova et al., 2021), or embedding an inverse mapping into a single network layer (Dong et al., 2021; Li et al., 2023). Such strategies incorporate physics in a one-sided manner, either operating purely in the measurement space or only after an inverse mapping. Yet blurry raw measurements contain information that is often lost once a direct inversion is applied, while using them alone leaves convolutional layers unable to reliably extract the underlying structure. This one-sided use of physics thus discards complementary cues that could stabilize reconstruction. To overcome this, architectures must sustain tight forward–inverse interactions throughout the hierarchy, so that both image-domain representations and raw measurement information are jointly leveraged within the physical system's constraints and remain aligned with real-world behavior.

To this end, we introduce **IFIN**, a unified encoder–decoder architecture that embeds differentiable forward operators and learnable inverse modules at every stage of the hierarchy. This design not only preserves physical consistency but also shapes richer feature representations by jointly leveraging measurement- and image-domain information. In addition, a physics-guided kernel adaptation compensates for imperfect or unknown PSF calibration; when direct PSF measurement is infeasible, the kernel is dynamically refined for blind deconvolution under the constraints of system physics. Our main contributions are as follows:

- **Integrated forward–inverse:** An encoder–decoder design where differentiable forward operators and learnable inverse modules are integrated at every level, maintaining physics consistency while shaping richer feature representations.
- **Learnable spatially variant modeling:** A parameterization that captures lateral shift- and depth-dependent variations, regularized for stability to enable accurate recovery in complex systems, with a learnable kernel representation jointly optimized with reconstruction, allowing blind deconvolution when PSF calibration is inaccurate or unavailable.

This explicit integration of forward and inverse processes is particularly effective in regimes where the measurements are heavily degraded, such as when the imaging system produces severely blurred data due to large point-spread functions. Data-driven inversion with conventional CNNs often struggles to capture the broader correlations required in these regimes, leaving much of the measurement signal underutilized. In contrast, our design propagates information through both the measurement and image domains at every stage, performing physics-guided inversion while simultaneously learning representations that capture variations difficult to model analytically. To illustrate the benefits of this physics-integrated framework, we focus on lensless imaging as a representative case, where our approach demonstrates superior performance compared to prior methods.

2 Related Work

2.1 Lensless Imaging

Lensless cameras replace conventional lenses with thin optical elements such as coded aperture, transmissive diffusers and engineered phase masks (Asif et al., 2016; Antipa et al., 2017; Lee et al., 2023b).As a result, diffuser- or mask-induced PSFs are large and highly structured, often encoding wide spatial neighborhoods—up to the entire scene—onto the sensor. This encoding eliminates the need for bulky optics but necessitates computational reconstruction to recover interpretable images from the raw measurements.

Beyond simple image recovery, the same computational framework also unlocks a wide range of modalities, including depth estimation (Antipa et al., 2017; Bagadthey et al., 2022), hyperspectral imaging (Sahoo et al., 2017; Monakhova et al., 2020), polarization analysis (Baek et al., 2022), single-shot ultrafast video capture via rolling-shutter coding (Antipa et al., 2019), and privacy-

preserving imaging based on the expressive representations of lensless measurements (Satat et al., 2017; Henry et al., 2023). These capabilities, coupled with the ultra-compact and lightweight architecture, make lensless cameras particularly appealing for applications in embedded vision where size, cost, and its unique imaging functionalities are critical (Kim et al., 2024; Ge et al., 2024; Xiangjun & Yue, 2025).

Image reconstruction becomes particularly challenging when the optical system produces strongly spread measurements, often modeled as 2-D or 3D convolutions with large kernels. In such cases, extended PSFs distribute scene information broadly across the sensor, leading to loss of spatial detail and strong overlap between measurements, which makes inversion ill-posed. Similar challenges arise in a range of computational imaging settings, from conventional cameras under severe aberrations or motion blur to advanced imaging tasks such as imaging through scattering media (Yoon et al., 2020), non-line-of-sight imaging (Faccio et al., 2020), coherent diffractive imaging (Miao et al., 2015) and advanced microscopy techniques with designed PSFs (Pavani et al., 2009).

We begin with a baseline shift-invariant model, which assumes that the system response is identical across all spatial locations. Under this assumption, the measurement is expressed as a 2D convolution between the scene and a position-independent PSF:

$$y[i,j] \;=\; \sum_{k,\ell} h[i-k,\, j-\ell]\, x[k,\ell] \;+\; \eta[i,j], \tag{1}$$

where $x, y \in \mathbb{R}^{H\times W}$ denote the scene irradiance and the captured measurement, $h[\cdot,\cdot]$ is a position-independent PSF, and η models additive noise.

In practice, most imaging systems are not truly shift-invariant. Off-axis aberrations (e.g., coma/astigmatism), depth-dependent propagation, field-dependent magnification, vignetting/pupil clipping, and sensor truncation at the image boundaries all make the effective system response depend on spatial location (Booth, 2014; Thiébaut et al., 2016; Antipa et al., 2017). This is especially pronounced when a phase or coded mask are non-planar or engineered for a high effective numerical aperture: resolution improves on-axis, but aberration-induced shift variance grows with field angle. As a result, the PSF $h_{i,j}$ widens, skews, or changes phase structure across the field, necessitating a spatially varying model. A more general shift-variant model accounts for this effect by allowing the PSF to vary with the output coordinates:

$$y[i,j] \;=\; \sum_{k,\ell} h_{i,j}[k,\ell]\, x[k,\ell] \;+\; \eta[i,j], \tag{2}$$

where $h_{i,j}[k,\ell]$ is a location-dependent PSF at pixel position $[k,\ell]$. This formulation no longer reduces to a simple 2D convolution, but instead to a large, spatially varying linear operator, which significantly increases computational and memory demands for inversion. Compounding the difficulty are sensor cropping (finite field-of-view truncation) and measurement noise, which complicate inversion and markedly increase ill-posedness and susceptibility to calibration errors.

2.2 Image Restoration

Given such forward models, image recovery in lensless cameras is carried out through computational inversion, often posed as deconvolution. This places lensless reconstruction in the same category of problems as deblurring in conventional cameras, where severe optical blur can arise from optical aberrations, motion, or atmospheric turbulence. A long line of approaches has been explored for this task. Classical methods—including Wiener deconvolution (Wiener, 1964), Richardson–Lucy (Richardson, 1972; Lucy, 1974) provide physically grounded solutions, but their performance quickly deteriorates under noise and kernel mis-specification. In practice, non-differentiable elements in the forward model (e.g., cropping or truncation) together with priors such as total variation (TV) regularization motivate the use of optimization frameworks like the Alternating Direction Method of Multipliers (ADMM) (Boyd et al., 2011), which decouple data fidelity and regularization terms to enable tractable iterative solvers (Antipa et al., 2017).

In parallel, substantial efforts address non-uniform blur using PSF fields—either calibrated or estimated (Robbins & Huang, 1972; Denis et al., 2015; Yeo et al., 2025), trading off additional calibration effort and computational cost. More recently, deep learning methods—including CNN-

based (Ronneberger et al., 2015; Zhang et al., 2018b; Anonymous, 2020; Chen et al., 2022) and vision-transformer (ViT) architectures (Dosovitskiy et al., 2020; Pan et al., 2022)—have emerged as powerful data-driven approaches, learning end-to-end mappings from measurements to images and often achieving state-of-the-art restoration quality, yet they remain prone to hallucinations and limited generalization under diverse real-world degradations.

Building on the limitations of purely data-driven approaches, a growing body of work has explicit physical models into neural reconstruction pipelines, rather than relying solely on a fully data-driven mapping from severely degraded measurements. One prominent direction unrolls classical optimization, embedding the forward operator directly into iterative updates: Monakhova et al. (2019) and Poudel & Nakarmi (2024) augment each iteration with a neural denoiser, while Kingshott et al. (2022) adopts a primal–dual unrolling that jointly learns forward and adjoint operators. Forward-model constraints are also used for measurement consistency in unsupervised training, where neural priors alone can guide reconstructions without ground-truth supervision (Ulyanov et al., 2018; Wang et al., 2020; Monakhova et al., 2021). Another approach employs feed-forward hybrid architectures, where a physics-based inversion stage is followed by a learned refinement network (Khan et al., 2020; Yanny et al., 2022). A related line of work performs deconvolution in feature space, embedding convolutional inversion within multiscale skip connections to improve fidelity and robustness (Dong et al., 2021; Li et al., 2023).

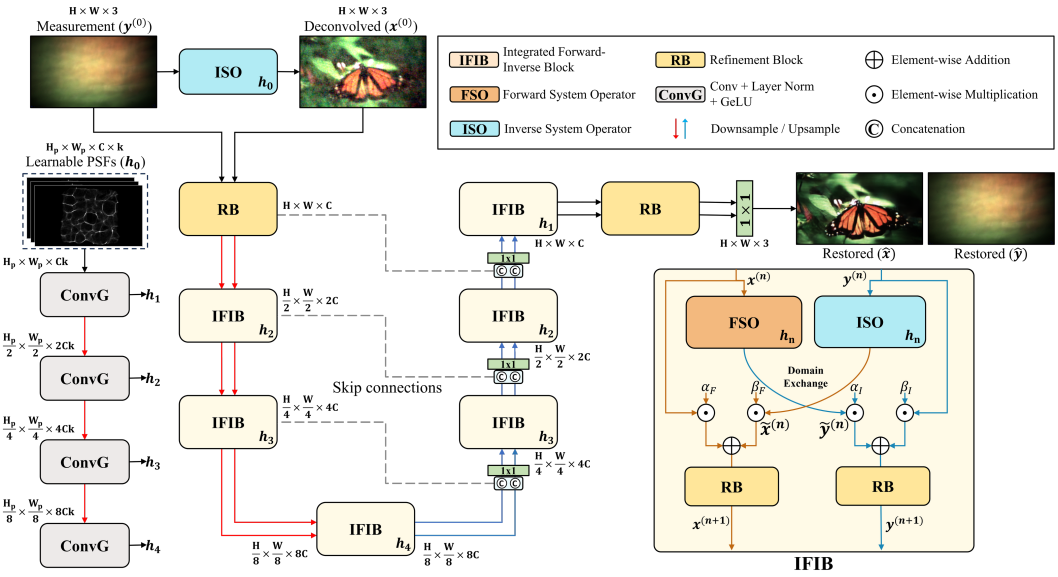

Figure 1: Overall architecture of IFIN. The network follows an encoder–decoder structure, where integrated Forward–Inverse Blocks (IFIBs) are inserted at each scale to jointly apply the Forward System Operator (FSO) and Inverse System Operator (ISO). A shared learnable PSF field guides both operators, ensuring forward–inverse consistency across scales.

Severe large-kernel blur mixes information across distant pixels, yet typical CNNs and ViT-style models process measurements within limited receptive fields or local windows, leaving much of this information unused (Luo et al., 2016; Liu et al., 2021). As a result, when raw measurements are fed directly into such architectures, their restoration ability becomes fundamentally constrained. Even when the physical inversion is embedded within neural layers (Dong et al., 2021; Li et al., 2023), the injected measurements collapse into degenerate feature representations, preventing optimal recovery. Moreover, one-sided physics–NN pipelines (Monakhova et al., 2019; Khan et al., 2020; Kingshott et al., 2022; Yanny et al., 2022; Poudel & Nakarmi, 2024) are limited in recovering components projected into the optical null space, often hallucinating such content from priors rather than reconstructing it deterministically. In this context, a restoration model should maintain the information contained in the raw measurements throughout the network while mitigating the null-space effects introduced by the inverse mapping.

3 METHOD

3.1 OVERVIEW

The proposed IFIN adopts an encoder–decoder backbone with Integrated Forward–Inverse Blocks (IFIBs) placed at every scale. As illustrated in Fig. 1, the encoder downsamples the input to capture large-scale blur effects and coarse scene structure, while the decoder progressively upsamples features to recover fine detail at the native resolution.

Feature-scale forward–inverse. We note that performing inverse operations in the feature space is also meaningful, as it facilitates the recovery of finer details during the network reconstruction process (Dong et al., 2021; Li et al., 2023). Conversely, embedding forward operations in the feature space encourages the learned representation to reflect the properties of the measurement.

PSF conditioning across scales. A learnable PSF field is first processed by a lightweight PSF encoder to produce multi-scale embeddings $\{h^{(n)}\}$. These embeddings condition both the Forward System Operator (FSO) (image→measurement) and the Inverse System Operator (ISO) (measurement→image) at the corresponding resolution, maintaining physical consistency across the hierarchy.

Initialization. We warm-start the reconstruction by applying the ISO to the raw measurement using the PSFs, producing a coarse inverse estimate. After a refinement, the pair *(measurement, coarse reconstruction)* enters the encoder–decoder as two coupled streams.

Scale-wise coupling. At each resolution stage, an IFIB jointly applies the ISO to the measurement stream and the FSO to the reconstruction stream. Features are exchanged bidirectionally between the two streams, so that measurement-domain consistency (via the FSO) and image-domain fidelity (via the ISO) are enforced in tandem rather than in a one-sided fashion.

3.2 LEARNABLE SPATIALLY VARYING PSFS

IFIN incorporates a learnable PSF representation that provides explicit system awareness to both FSO and ISO. The PSF field is parameterized as $k{=}s^2$ kernels covering local regions of the image. In case of $s{=}1$ (i.e., $k{=}1$), the PSF field reduces to a single global kernel. Kernels are initialized from calibrated measurements, a single reference PSF, or random patterns, and are optimized jointly with the network. A compact PSF encoder maps the field to multi-scale embeddings $\{h_n\}$ that condition all IFIBs throughout the network. By embedding the learnable PSF field, IFIN adapts to unknown or mismatched degradations and supports blind kernel estimation without external calibration, helping both operators remain physically consistent even under severe, spatially varying blur.

3.3 INTEGRATED FORWARD–INVERSE BLOCK (IFIB)

The **IFIB** is the fundamental unit of IFIN, designed to couple forward and inverse imaging processes at each scale. Each IFIB consists of two parallel operators: (1) a Forward System Operator (FSO), and (2) an Inverse System Operator (ISO), as illustrated in Fig. 2. Both operators are fundamentally tied to the target system's forward and inverse physics. In practice, however, they can be flexibly configured: a purely shift-invariant model can be used when degradations are approximately uniform, or a spatially varying operator can be invoked to handle more complex degradations. This adaptability allows IFIN to balance efficiency and fidelity across different imaging conditions.

3.3.1 FORWARD SYSTEM OPERATOR (FSO)

FSO simulates how the current estimate $\hat{x}$ would be formed by the physical system. By default we use 2-D linear 2-D linear convolution with zero padding via a single point-spread function h:

$$\tilde{y}[i,j] \;=\; (\hat{x} * h)[i,j]. \tag{3}$$

When nonstationarity is present, we can model the FSO as a fully shift-variant convolution by tiling $\hat{x}$ and applying local PSFs with normalized overlap–add reassembly (See Appendix A.10). For computational efficiency in IFIN, we use a single-convolution surrogate with the averaged PSF

$h_{\text{eff}} = \frac{1}{k}\sum_{r=1}^{k} h_r$ while preserving the measurement-consistency signal. We perform a convolution in Eq. (3) with h_{eff} instead of h.

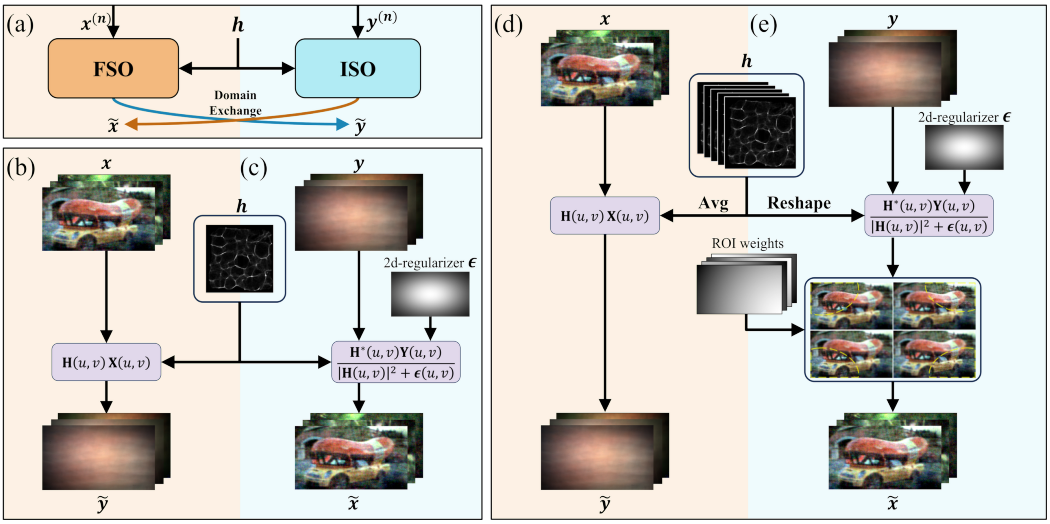

Figure 2: (a) Illustration of the forward–inverse pair inside an IFIB—the submodule that hosts the Forward System Operator (FSO) and the Inverse System Operator (ISO); (b) FSO under shift-invariant condition; (c) ISO under shift-invariant condition; (d) FSO under shift-variant condition; (e) ISO under shift-variant condition.

3.3.2 Inverse System Operator (ISO)

ISO restores a sharp estimate from the degraded measurement via Wiener-like deconvolution with a learnable frequency-dependent regularizer. By default, for PSF h, letting $Y(u,v) = \mathcal{F}\{\, W P_{rp} y \,\}(u,v)$ and $H(u,v) = \mathcal{F}\{h\}(u,v)$, where $\mathcal{F}\{\cdot\}$ denotes the Fourier transform, P_{rp} is replicate padding, and W is a mild Gaussian window used to mitigate wrap-around artifacts during deconvolution (Khan et al., 2020), we compute:

$$\widehat{X}(u,v) \;=\; \frac{H^{*}(u,v)}{|H(u,v)|^{2} + \epsilon(u,v)}\, Y(u,v), \qquad \epsilon(u,v) \geq 0, \tag{4}$$

and set $\hat{x} = \mathcal{F}^{-1}\{\widehat{X}\}$. Here, $\epsilon(u,v)$ is a 2-D learnable parameterization predicted and refined during training, with nonnegativity enforced by a ReLU. By learning to estimate the distribution of noise variance from data, ϵ functions as an optimal frequency-selective prior that adapts to system noise.

In the spatially varying case, we perform region-wise Wiener-like deconvolution with distinct PSFs and blend the partial reconstructions:

$$Y(u,v) \;=\; \mathcal{F}\big\{\, W\, P_{rp}\big(y\big)\big\}(u,v), \qquad H_r(u,v) = \mathcal{F}\{h_r\}(u,v), \tag{5}$$

$$\widehat{X}_r(u,v) \;=\; \frac{H_r^{*}(u,v)}{|H_r(u,v)|^{2} + \epsilon_r(u,v)}\, Y(u,v), \qquad \epsilon_r(u,v) \geq 0 \text{ (ReLU)}, \tag{6}$$

$$\hat{x}[i,j] \;=\; \sum_{r=1}^{m} w_r[i,j]\, \mathcal{F}^{-1}\{\widehat{X}_r\}[i,j], \tag{7}$$

where $\{w_r\}_{r=1}^{m}$ are learnable region-of-interest (ROI) maps, allowing the model to optimize the spatial support of each PSF over regions of varying extent and location. We initialize the ROI maps from Gaussian kernels $\{g_r\}_{r=1}^{k}$ centered at $\{p_r\}$, where $\{p_r\}$ are the centers of an $s \times s$ grid partitioning the input measurement.

$$g_r[i,j] \;=\; \exp\!\Big(-\tfrac{\|(i,j)-p_r\|_2^2}{2\sigma_r^2}\Big), \qquad w_r[i,j] \;=\; \frac{g_r[i,j]}{\sum_{q=1}^{m} g_q[i,j]}, \qquad \sum_{r=1}^{m} w_r[i,j] \;=\; 1 \;\; \forall (i,j). \tag{8}$$

This construction explicitly recovers different spatial neighborhoods using region-specific PSFs and region-specific frequency priors, which is critical under spatially varying blur.

3.3.3 INTEGRATED FORWARD-INVERSE

The hallmark of IFIB is the bidirectional exchange between FSO and ISO. Each operator contributes complementary information to the other: FSO enforces measurement-domain consistency, while ISO sharpens image-domain details. We implement this by passing the output of each operator as a skip connection into the input of its counterpart. This ensures both branches evolve jointly:

$$y^{(n+1)} = \phi_\theta^y\left(\alpha^{(n)} \cdot y^{(n)} + \beta^{(n)} \cdot \tilde{y}^{(n)}\right), \qquad x^{(n+1)} = \phi_\theta^x\left(\alpha^{(n)} \cdot x^{(n)} + \beta^{(n)} \cdot \tilde{x}^{(n)}\right). \tag{9}$$

where ϕ_θ^y and ϕ_θ^x are lightweight refinement modules, and $\alpha^{(n)}$ and $\beta^{(n)}$ are learned scalar gates at scale n, with n indexing the IFIB stage within the encoder–decoder hierarchy.

Refinement block (RB) To boost performance, we adopt a refinement block that applies learned priors to stabilize and refine the physics-transformed features, enabling coarse-to-fine reconstruction. Following Chen et al. (2022), the RB is a normalization-free residual module built from depthwise-separable convolutions with a simple channel-gating mechanism. We place lightweight convolutional layers before and after the core to stabilize the feature statistics produced by the forward–inverse integrated.

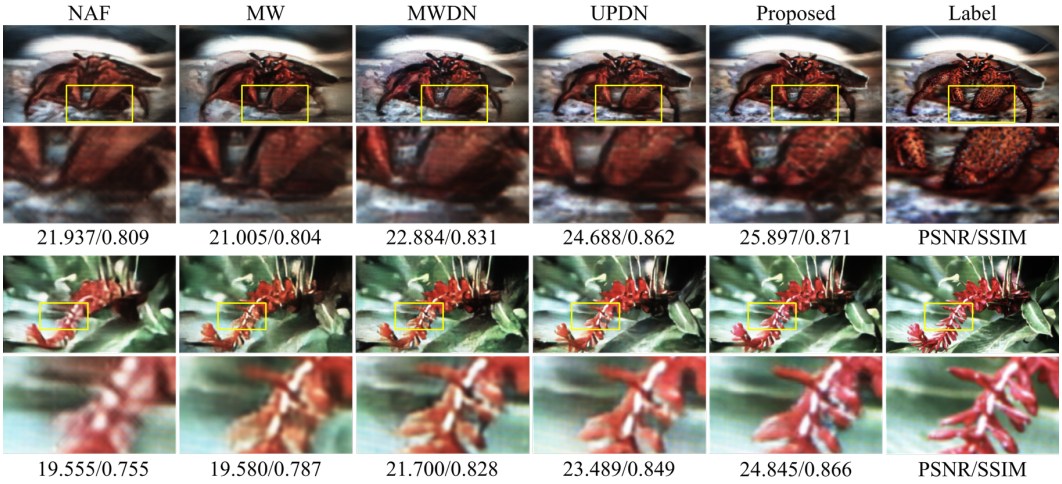

Figure 3: Qualitative comparison on DiffuserCam dataset.

4 EXPERIMENTS

To evaluate IFIN, we conduct end-to-end supervised training on large sets of scene–measurement pairs, including both real display-and-capture data and synthetic degradations. Training pairs are constructed through three routes:

DiffuserCam (Monakhova et al., 2019)— display-and-capture measurements acquired with a diffuser-based lensless camera, using co-located reference camera images as ground-truth labels;

Custom Shift-Variant (SV) Lensless—display-and-capture using our high-resolution, wide-field phase-mask-based lensless camera, aligned directly to the original display images; and

MultiWienerNet (MW) (Yanny et al., 2022)—synthetic training pairs generated by convolving ground-truth images with spatially variant PSFs measured from a mask-based microscope (miniscope), with validation performed on real miniscope captures.

Further details on dataset composition, camera assembly, registration procedures, and a summary of the experimental setups and evaluation protocols are provided in the Supplementary Material (Appendix A.3 and Appendix A.4).

We evaluate our method against traditional approaches (Wiener deconvolution, ADMM-TV) and learning-based models, including data-driven approaches (U-Net, NAFNet) and physics-guided approaches (Le-ADMM-U, DeepLIR, MWNet, UPDN, and MWDN). Details of the baselines and training protocols are provided in Appendix A.2.

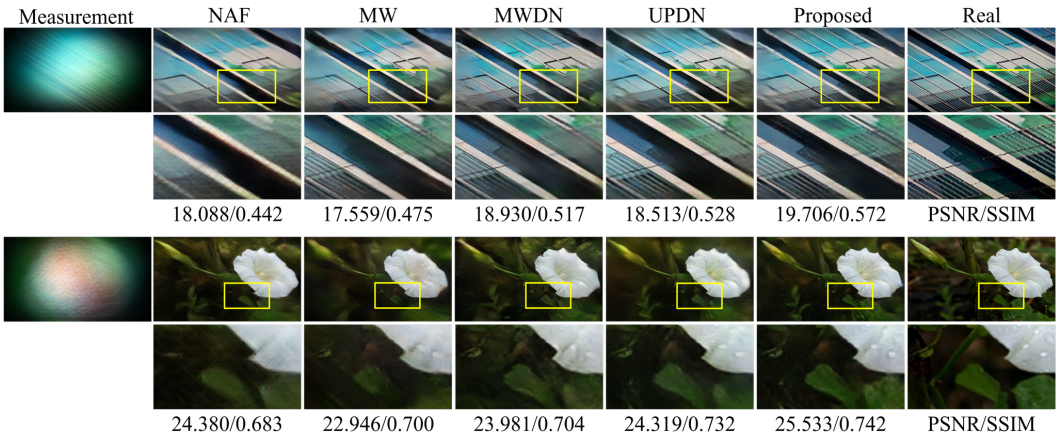

Figure 4: Qualitative comparison on our proposed SV Lensless dataset.

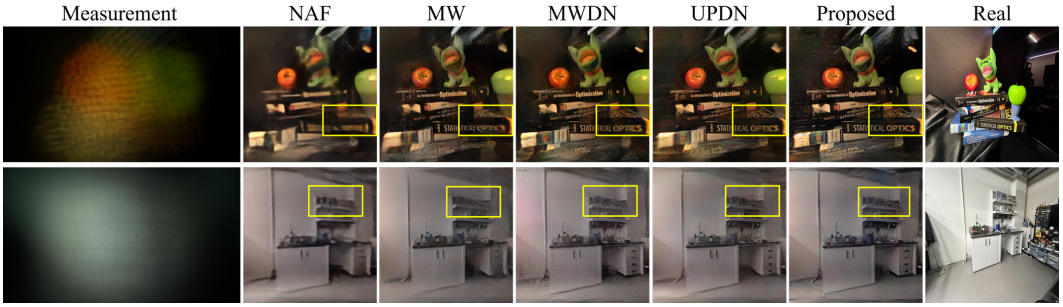

Figure 5: Reconstruction of in-the-wild lenless imaging. IFIN generalizes robustly beyond controlled settings, recovering fine features and natural textures from real-world lensless measurements.

Table 1: Quantitative comparison on three benchmarks—**DiffuserCam**, **SV Lensless**, and **MultiWienerNet**. We report PSNR ↑, LPIPS ↓ (Zhang et al., 2018a), and SSIM ↑ (arrows indicate the preferred direction). Classical baselines (ADMM, Wiener Deconvolution), a single **ISO** model, and recent learning-based methods are included.

Dataset	DiffuserCam			SV Lensless			MultiWienerNet		
Metrics	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑
ADMM	12.252	0.607	0.346	11.843	0.643	0.323	19.189	0.557	0.420
Wiener Deconv.	12.552	0.591	0.384	12.405	0.607	0.369	18.658	0.640	0.302
ISO	16.528	0.544	0.404	17.240	0.462	0.444	20.202	0.623	0.380
UNet	21.230	0.394	0.656	21.890	0.474	0.646	23.859	0.389	0.589
NAFNet	24.830	0.239	0.810	23.857	0.245	0.769	24.657	0.282	0.712
Le-ADMM-U	23.261	0.312	0.765	21.956	0.278	0.748	23.732	0.335	0.702
DeepLIR	25.958	0.260	0.829	20.523	0.339	0.642	22.556	0.379	0.642
MWNet	24.832	0.247	0.810	23.001	0.255	0.766	25.660	0.260	0.728
UPDN	28.228	0.194	0.877	23.920	0.229	0.801	24.364	0.287	0.707
MWDNet-CPSF	27.298	0.217	0.845	24.525	0.224	0.801	27.436	0.236	0.780
IFIN (Ours)	**29.730**	**0.180**	**0.888**	**25.444**	**0.201**	**0.824**	**31.083**	**0.175**	**0.866**

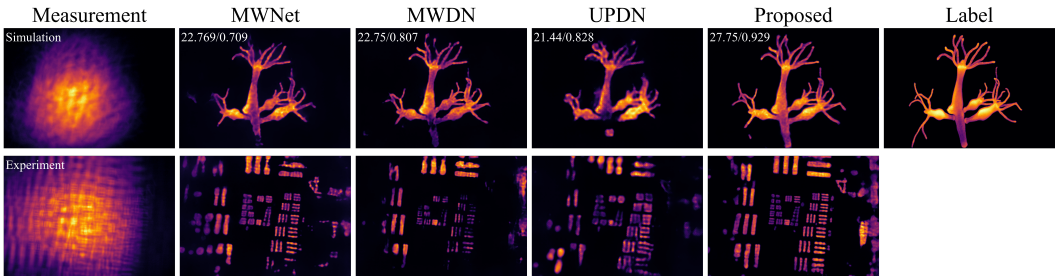

Figure 6: Qualitative comparison on the MultiWienerNet dataset, with simulated data with spatially variant PSFs and a real miniscope capture of USAF resolution targets.

4.1 Results

DiffuserCam. Figure 3 illustrates reconstruction results on display–capture data. Our method delivers accurate color reproduction and sharper reconstruction of fine structures, surpassing prior approaches that frequently suppress details or introduce artifacts. The advantages are especially clear near edges and in low-contrast regions, where competing methods typically fail to preserve mid-frequency detail and color fidelity. Consistent with these observations, our method also achieves the highest quantitative scores among all baselines, as summarized in Table 1.

SV Lensless. Our in-house dataset presents two distinct challenges. First, its wide field of view leads to strong PSF shift-variance in the outer regions, where conventional shift-invariant models fail and produce heavy blur and distortion. Second, our custom mask provides higher intrinsic resolution with sharper PSFs, so the reconstruction must preserve fine structures and deliver correspondingly sharp features. Our method succeeds on both fronts: it reconstructs central content cleanly while maintaining detail and contrast in the periphery through explicit shift-variant modeling and feature-level deconvolution, without requiring direct measurement of off-axis PSFs. As shown in Figure 4, the results exhibit sharper details and fewer distortions across the field compared to competing approaches. Figure 5 shows real lensless captures without ground truth. **IFIN** demonstrates strong generalization to real-world degradations.

MultiWienerNet. On the MW dataset, our method yields the sharpest and most faithful reconstructions in both simulated and real captures (Figure 6). In simulated data, IFIN closely matches the ground-truth structure and contrast, attaining the best PSNR/SSIM. On real miniscope measurements, it preserves the widest effective field of view and the highest apparent acuity on USAF targets, while suppressing ringing and better separating closely spaced line pairs than competing methods. Notably, MWNet reconstructs with 9-*calibrated* PSFs, which explains its advantage over approaches that rely on a single on-axis PSF. In contrast, IFIN is initialized only with an on-axis PSF, while *learning* a field-dependent inverse by estimating effective PSFs across the FoV during training (see Appendix A.7 for details on PSF estimation). This enables the model to handle shift variance without dense per-field calibration. This learning is coupled with feature-level deconvolution, which further improves fidelity and robustness under calibration mismatch and measurement noise. As a result, whereas MWDN (Li et al., 2023) remains strong in simulation but degrades on experimental data, IFIN adapts more reliably to the real forward model.

5 Conclusion

We presented **IFIN**, a physics-guided reconstruction network that interleaves forward and inverse operators at every scale with a learnable PSF field for calibration-free, spatially varying awareness. Experiments and ablations show that tight forward–inverse coupling and adaptive PSF modeling are crucial for high-fidelity recovery under large-support, shift-variant kernels. We expect IFIN to serve as a general template for embedding physical operators into learned reconstructions across various computational imaging systems.

6 Ethics Statement

This study presents an advanced reconstruction method for inverse problems with potential relevance to domains such as medical imaging, security inspection, and optical system design. We emphasize responsible research practices by carefully considering ethical implications. Our experiments do not involve human subjects, personal data, or sensitive information; instead, the datasets are either synthetically generated or collected under controlled settings to avoid privacy risks. While the proposed method could in principle be applied to biomedical or security-related contexts, the present work is intended solely for scientific progress. Any practical deployment in sensitive areas must comply with applicable ethical standards and obtain proper regulatory approvals. Furthermore, we consider and address potential biases in both data and models to promote fairness, robustness, and broad generalizability across diverse imaging conditions. By following these principles, our research aims to contribute responsibly to computational imaging and the broader study of reconstruction.

References

Anonymous. Details omitted for double-blind review, 2020.

Anonymous. Details omitted for double-blind review, 2023.

Nick Antipa, Grace Kuo, Reinhard Heckel, Ben Mildenhall, Emrah Bostan, Ren Ng, and Laura Waller. Diffusercam: lensless single-exposure 3d imaging. *Optica*, 5(1):1–9, 2017.

Nick Antipa, Patrick Oare, Emrah Bostan, Ren Ng, and Laura Waller. Video from stills: Lensless imaging with rolling shutter. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–8. IEEE, 2019.

M Salman Asif, Ali Ayremlou, Aswin Sankaranarayanan, Ashok Veeraraghavan, and Richard G Baraniuk. Flatcam: Thin, lensless cameras using coded aperture and computation. *IEEE Transactions on Computational Imaging*, 3(3):384–397, 2016.

Nakkyu Baek, Yujin Lee, Taeyoung Kim, Jaewoo Jung, and Seung Ah Lee. Lensless polarization camera for single-shot full-stokes imaging. *APL Photonics*, 7(11), 2022.

Dhruvjyoti Bagadthey, Sanjana Prabhu, Salman S Khan, D Tony Fredrick, Vivek Boominathan, Ashok Veeraraghavan, and Kaushik Mitra. Flatnet3d: intensity and absolute depth from single-shot lensless capture. *Journal of the Optical Society of America A*, 39(10):1903–1912, 2022.

Martin J Booth. Adaptive optical microscopy: the ongoing quest for a perfect image. *Light: Science & Applications*, 3(4):e165–e165, 2014.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, 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.

Xin Cai, Zhiyuan You, Hailong Zhang, Jinwei Gu, Wentao Liu, and Tianfan Xue. Phocolens: Photorealistic and consistent reconstruction in lensless imaging. *Advances in Neural Information Processing Systems*, 37:12219–12242, 2024.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pp. 17–33. Springer, 2022.

Loïc Denis, Eric Thiébaut, Ferréol Soulez, Jean-Marie Becker, and Rahul Mourya. Fast approximations of shift-variant blur. *International Journal of Computer Vision*, 115(3):253–278, 2015.

Jiangxin Dong, Stefan Roth, and Bernt Schiele. Dwdn: Deep wiener deconvolution network for non-blind image deblurring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9960–9976, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Daniele Faccio, Andreas Velten, and Gordon Wetzstein. Non-line-of-sight imaging. *Nature Reviews Physics*, 2(6):318–327, 2020.

Haoyang Ge, Qiao Feng, Hailong Jia, Xiongzheng Li, Xiangjun Yin, You Zhou, Jingyu Yang, and Kun Li. Lpsnet: End-to-end human pose and shape estimation with lensless imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1471–1480, 2024.

Chris Henry, M Salman Asif, and Zhu Li. Privacy preserving face recognition with lensless camera. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pp. 39–43, 2008.

Salman Siddique Khan, Varun Sundar, Vivek Boominathan, Ashok Veeraraghavan, and Kaushik Mitra. Flatnet: Towards photorealistic scene reconstruction from lensless measurements. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):1934–1948, 2020.

Taeyoung Kim, Kyung Chul Lee, Kyungwon Lee, Nakkyu Baek, Jaewoo Jung, Eosu Kim, Bobae Park, Junghee Ha, Keun You Kim, Young-Seok Seo, et al. High-speed lensless eye tracker for microsaccade measurement. In *SPIE Advanced Biophotonics Conference (SPIE ABC 2023)*, volume 13076, pp. 38–44. SPIE, 2024.

Oliver Kingshott, Nick Antipa, Emrah Bostan, and Kaan Akşit. Unrolled primal-dual networks for lensless cameras. *Optics Express*, 30(26):46324–46335, 2022.

Grace Kuo, Fanglin Linda Liu, Irene Grossrubatscher, Ren Ng, and Laura Waller. On-chip fluorescence microscopy with a random microlens diffuser. *Optics express*, 28(6):8384–8399, 2020.

Chanseok Lee, Gookho Song, Hyeonggeon Kim, Jong Chul Ye, and Mooseok Jang. Deep learning based on parameterized physical forward model for adaptive holographic imaging with unpaired data. *Nature Machine Intelligence*, 5(1):35–45, 2023a.

Kyung Chul Lee, Junghyun Bae, Nakkyu Baek, Jaewoo Jung, Wook Park, and Seung Ah Lee. Design and single-shot fabrication of lensless cameras with arbitrary point spread functions. *Optica*, 10(1):72–80, 2023b.

Ying Li, Zhengdai Li, Kaiyu Chen, Youming Guo, and Changhui Rao. Mwdns: reconstruction in multi-scale feature spaces for lensless imaging. *Optics Express*, 31(23):39088–39101, 2023.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

Leon B Lucy. An iterative technique for the rectification of observed distributions. *Astronomical Journal, Vol. 79, p. 745 (1974)*, 79:745, 1974.

Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *NeurIPS*, 2016.

Jianwei Miao, Tetsuya Ishikawa, Ian K Robinson, and Margaret M Murnane. Beyond crystallography: Diffractive imaging using coherent x-ray light sources. *Science*, 348(6234):530–535, 2015.

Kristina Monakhova, Joshua Yurtsever, Grace Kuo, Nick Antipa, Kyrollos Yanny, and Laura Waller. Learned reconstructions for practical mask-based lensless imaging. *Optics express*, 27(20): 28075–28090, 2019.

Kristina Monakhova, Kyrollos Yanny, Neerja Aggarwal, and Laura Waller. Spectral diffusercam: lensless snapshot hyperspectral imaging with a spectral filter array. *Optica*, 7(10):1298–1307, 2020.

Kristina Monakhova, Vi Tran, Grace Kuo, and Laura Waller. Untrained networks for compressive lensless photography. *Optics Express*, 29(13):20913–20929, 2021.

Xiuxi Pan, Xiao Chen, Saori Takeyama, and Masahiro Yamaguchi. Image reconstruction with transformer for mask-based lensless imaging. *Optics Letters*, 47(7):1843–1846, 2022.

Sri Rama Prasanna Pavani, Michael A Thompson, Julie S Biteen, Samuel J Lord, Na Liu, Robert J Twieg, Rafael Piestun, and William E Moerner. Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function. *Proceedings of the National Academy of Sciences*, 106(9):2995–2999, 2009.

Arpan Poudel and Ukash Nakarmi. 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.

William Hadley Richardson. Bayesian-based iterative method of image restoration. *Journal of the optical society of America*, 62(1):55–59, 1972.

Gregory M Robbins and Thomas S Huang. Inverse filtering for linear shift-variant imaging systems. *Proceedings of the IEEE*, 60(7):862–872, 1972.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Sujit Kumar Sahoo, Dongliang Tang, and Cuong Dang. Single-shot multispectral imaging with a monochromatic camera. *Optica*, 4(10):1209–1213, 2017.

Guy Satat, Matthew Tancik, Otkrist Gupta, Barmak Heshmat, and Ramesh Raskar. Object classification through scattering media with deep learning on time resolved measurement. *Optics express*, 25(15):17466–17479, 2017.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8922–8931, 2021.

Éric Thiébaut, Loïc Dénis, Ferréol Soulez, and Rahul Mourya. Spatially variant psf modeling and image deblurring. In *Adaptive Optics Systems V*, volume 9909, pp. 2211–2220. SPIE, 2016.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9446–9454, 2018.

Fei Wang, Yaoming Bian, Haichao Wang, Meng Lyu, Giancarlo Pedrini, Wolfgang Osten, George Barbastathis, and Guohai Situ. Phase imaging with an untrained neural network. *Light: Science & Applications*, 9(1):77, 2020.

Norbert Wiener. *Extrapolation, interpolation, and smoothing of stationary time series*. The MIT press, 1964.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16133–16142, 2023.

Yin Xiangjun and Huihui Yue. Reveal object in lensless photography via region gaze and amplification. In *The Thirteenth International Conference on Learning Representations*, 2025.

Kyrollos Yanny, Nick Antipa, William Liberti, Sam Dehaeck, Kristina Monakhova, Fanglin Linda Liu, Konlin Shen, Ren Ng, and Laura Waller. Miniscope3d: optimized single-shot miniature 3d fluorescence microscopy. *Light: Science & Applications*, 9(1):171, 2020.

Kyrollos Yanny, Kristina Monakhova, Richard W Shuai, and Laura Waller. Deep learning for fast spatially varying deconvolution. *Optica*, 9(1):96–99, 2022.

Joel Yeo, N Duane Loh, Ramon Paniagua-Dominguez, and Arseniy I Kuznetsov. Eigencwd: a spatially varying deconvolution algorithm for single metalens imaging. *Optics Express*, 33(13): 28481–28492, 2025.

Seokchan Yoon, Moonseok Kim, Mooseok Jang, Youngwoon Choi, Wonjun Choi, Sungsam Kang, and Wonshik Choi. Deep optical imaging within complex scattering media. *Nature Reviews Physics*, 2(3):141–158, 2020.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 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, 2018a.

Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 286–301, 2018b.

A Appendix

A.1 Training Details

We train IFIN using the AdamW optimizer with a learning rate of 1×10^{-4} and $(\beta_1, \beta_2) = (0.9, 0.999)$ for all parameters except the point spread function (PSF). For the PSF, we employ a separate AdamW optimizer with a learning rate of 1×10^{-3}. Both learning rates are reduced by a factor of 0.5 when the validation loss plateaus. The training is conducted with a batch size of 4.

To initialize the network, for spatially varying deconvolution we construct ROI maps aligned to the input size, enabling local adaptation without excessive computational overhead. In the forward–inverse integration, the weighting parameters are initialized as $\alpha = 0.8$ and $\beta = 0.2$, balancing the contributions of the forward and inverse operators at the start of training.

For the PSF representation, to reduce computational cost we crop the support region of the on-axis PSF to include only its effective feature area. Under shift-variant conditions, we do not rely on calibrated PSFs; instead, the on-axis PSF is either replicated k times or randomly initialized. We experiment with $k = \{1, 4, 9, 16\}$ on DiffuserCam, and set $k = 9$ for both SV Lensless and MW datasets.

To ensure fairness, we match hyperparameters to those used in prior works. Whenever a network requires PSF inputs, we apply proper normalization (e.g., ℓ_2, ℓ_1, or max normalization) for stable training and meaningful results. Dataset-specific preprocessing steps such as cropping or affine registration are not incorporated into the loss, but only applied for visualization. Specifically, DiffuserCam images (480×270) are cropped to 360×210 (TBC); SV Lensless inputs (480×270) are aligned by applying the inverse of the affine transform used in label registration, with bicubic interpolation; and MW dataset images are resized from 640×448 to 320×224 to reduce computational cost. Further details on dataset preparation and affine transforms are provided in A-2 and A-3.

Training IFIN with $k = 9$ at the resolution of DiffuserCam and SV Lensless datasets takes approximately 64 hours on a single NVIDIA A6000 GPU.

Loss Functions. We minimize a composite objective balancing pixel fidelity, perceptual quality, cross-domain consistency, and a physics prior on the PSF. Let y be the observed input, x the ground truth, $\hat{x}$ the final reconstructed output, $\hat{y}$ an intermediate output indicated as measurement, h the learned PSF, $\tilde{x}^{(0)}$ the first ISO-branch output, and $\mathcal{I}(\cdot, h)$ the ISO operator with PSF h.

$$\mathcal{L}_{\text{img}} = \|\hat{x} - x\|_2^2 \qquad \text{fidelity loss} \tag{10}$$

$$\mathcal{L}_{\text{perc}} = \text{LPIPS}_{\text{VGG}}(\hat{x}, x) \qquad \text{perceptual loss} \tag{11}$$

$$\mathcal{L}_{ISO} = \|\tilde{x}^{(0)} - x\|_2^2 \qquad \text{ISO supervision} \tag{12}$$

$$\mathcal{L}_{\text{cons}_{\text{im}}} = \|\hat{y} - y\|_2^2 \qquad \text{measurement consistency} \tag{13}$$

$$\mathcal{L}_{\text{cons}_{\text{ft}}} = \|\mathcal{S}(\hat{y}, h) - \hat{x}\|_2^2 \qquad \text{cross-domain alignment} \tag{14}$$

$$\mathcal{L}_{\text{psf}} = \|\min(h, 0)\|_1 \qquad \text{PSF non-negativity} \tag{15}$$

$$\mathcal{L} = \lambda_{\text{img}}\mathcal{L}_{\text{img}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} + \lambda_w\mathcal{L}_w + \lambda_{\text{cons}_{\text{im}}}\mathcal{L}_{\text{cons}_{\text{im}}} + \lambda_{\text{cons}_{\text{ft}}}\mathcal{L}_{\text{cons}_{\text{ft}}} + \lambda_{\text{psf}}\mathcal{L}_{\text{psf}} \tag{16}$$

$$(\lambda_{\text{img}},\, \lambda_{\text{perc}},\, \lambda_{w},\, \lambda_{\text{cons}_{\text{im}}},\, \lambda_{\text{cons}_{\text{ft}}},\, \lambda_{\text{psf}}) = (1.0,\, 0.05,\, 0.1,\, 0.01,\, 0.01,\, 0.1). \tag{17}$$

Coefficients are selected by validation and fixed across all experiments. For fairness in the comparison tables, baseline models retrained under our pipeline were optimized only with the image and perceptual losses ($\mathcal{L}_{\text{img}}$ and $\mathcal{L}_{\text{perc}}$) using consistent weights across datasets.

A.2 Baselines

We summarize each baseline we use with the reference, its implication in our study.

Wiener Deconvolution (Wiener, 1964): Classical closed-form deconvolution using a calibrated PSF. We tune the noise–to–signal parameter on a validation split. It is fast and simple but sensitive to noise and kernel mis-specification.

Alternating Direction Method of Multipliers (ADMM–TV) (Boyd et al., 2011): Variational reconstruction with a total-variation prior solved via ADMM, which decouples data fidelity and regularization; we use shared stopping criteria across scenes.

U-Net (Ronneberger et al., 2015): A simple encoder–decoder CNN with skip connections that serves as a purely data-driven reconstruction baseline.

NAFNet (Chen et al., 2022): A modern, parameter-efficient CNN restorer built from *Nonlinear Activation Free* (NAF) blocks, which replace explicit nonlinear activations with lightweight gating and normalization. As a recent CNN architecture for image restoration, NAFNet provides a strong purely data-driven baseline and helps illustrate how a modern denoiser design compares with physics-driven or hybrid methods.

Learned-ADMM-U (LE-ADMM-U) (Monakhova et al., 2019): Unrolled ADMM with learnable proximal operators over K iterations and U-Net denoiser. Its simple structure proposes a combination of physics and neural networks.

DeepLIR (Poudel & Nakarmi, 2024): Learned iterative reconstruction that updates with a learned denoiser over K steps; The study utilizes the ConvNeXt blocks (Woo et al., 2023) in the denoiser for strong attention-based approach in convolutional layers.

Unrolled Primal–Dual Network (UPDN) (Kingshott et al., 2022): Primal–dual unrolling with learnable update operators and step sizes and a final denoising stage is applied.

MultiWienerNet (MWNet) (Yanny et al., 2022): A lightweight physics-aware baseline that linearly combines multiple Wiener-filter outputs and refines them with U-Net under spatially varying conditions. When calibration is unavailable, we use an on-axis PSF and instantiate k filters accordingly for this baseline.

MultiWiener Deconvolution Network (MWDN) (Li et al., 2023): Feature-space deconvolution within a multi-scale encoder–decoder: Wiener-like inversions are inserted along skip connections to improve fidelity and robustness under severe blur.

A.3 Detailed Description of Public Datasets

DiffuserCam. The DiffuserCam dataset (Monakhova et al., 2019) comprises 25,000 paired captures acquired simultaneously with a mask-based lensless camera (Antipa et al., 2017) and a reference lensed camera aligned via a beam splitter, using images from MIRFlickr (Huiskes & Lew, 2008) displayed on a computer monitor. DiffuserCam prototype consists of an off-the-shelf Light Shaping Diffuser (Luminit $0.5°$) with a laser-cut paper aperture, the assembly positioned approximately

$9\,\mathrm{mm}$ in front of the sensor plane. Raw sensor frames of 1080×1920 pixels are downsampled by a factor of 4 to 270×480. The split uses 24,000 images for training and 1,000 for testing. A single PSF is calibrated at the field center using an on-axis LED point source at the screen plane.

MultiWienerNet. Built on microscope data from Miniscope3D (Yanny et al., 2020), the MultiWienerNet dataset (Yanny et al., 2022) explicitly calibrates spatially varying PSFs across the field: a sub-resolution bead is scanned to measure the PSF at multiple sensor locations, effectively sampling a 3×3 grid over the imaging field. Using these measured PSFs, a synthetic training set is generated by convolving natural images with the spatially varying forward model and adding Poisson and Gaussian noise to emulate realistic measurements. This yields 22,125 two-dimensional paired samples, split 80/20 for training/testing. All training data are simulated at the system's sensor field-of-view resolution, and the trained model is finally evaluated on real lensless measurements from the calibrated setup (Yanny et al., 2022). In this work, we resize the images to 320×224, considering computational cost.

A.4 SV Lensless Dataset

We introduce a new dataset built around a multi–lens-array–like phase mask engineered for a compact, high-resolution lensless camera. The mask profile is optimized for our optical design and fabricated via grayscale lithography. Using a Sony IMX708 sensor, we assemble the camera by placing a $2\,\mathrm{mm}$ aperture and the phase mask at approximately $1.6\,\mathrm{mm}$ from the sensor. By design, we prioritize on-axis resolution at the cost of increased PSF shift-variance with incident angle. We capture 4608×2592 measurements from sensor and resize them to 480×270. MIRFlickr images are displayed on an OLED at a working distance of $30\,\mathrm{cm}$ for acquisition as shown in Figure 7. For target alignment, we first reconstruct with a deconvolution baseline, estimate an affine transform between the reconstruction and the label, and apply this transform to the label images during training as shown in Figure 8.

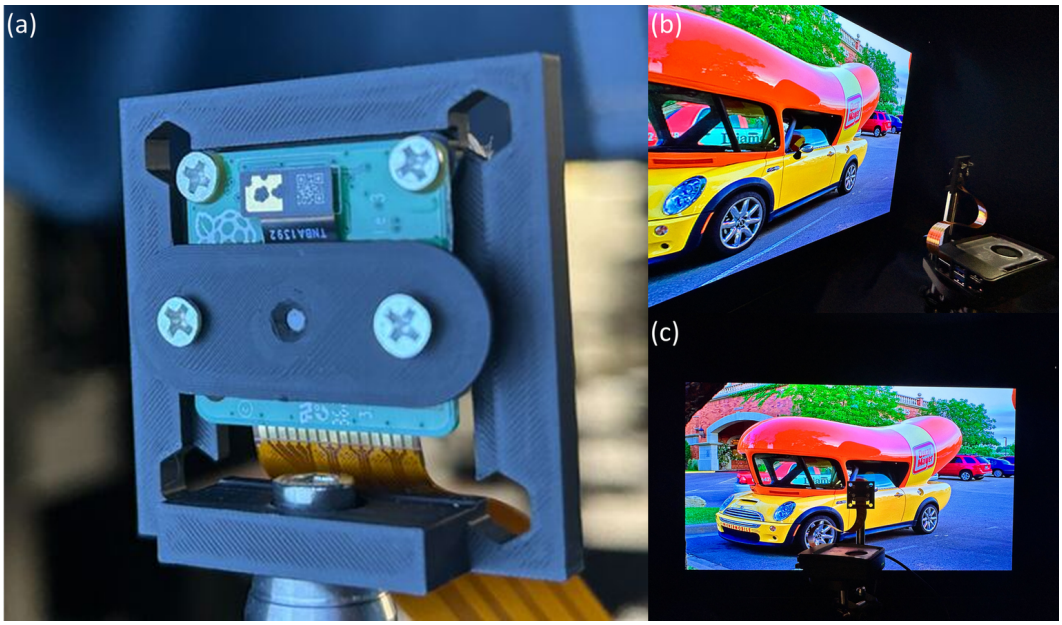

Figure 7: **Prototype lensless camera and dataset capture setup.** (a) Lensless camera prototype with a CMOS sensor mounted in a custom 3D-printed holder. (b–c) Display-and-capture configuration used for the SV Lensless dataset: reference images are rendered on a calibrated display while the prototype records the corresponding lensless measurements at a fixed geometry.

Optical Design Constraints

We designed the phase mask in a deductive manner under the mechanical and optical constraints. The constraints included: (i) RPI3 sensor module's mechanical stack and housing, (ii) target field of

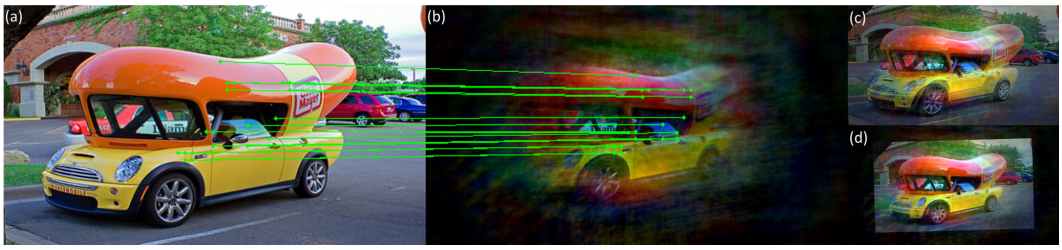

Figure 8: **Affine registration for label and display-capture pairs.** (a) Reference image shown on the display. (b) Raw lensless measurement captured by the prototype; green segments indicate LoFTR-based feature correspondences used to estimate the affine transform (Sun et al., 2021). (c) Overlay before registration. (d) Overlay after applying the estimated affine warp, yielding pixel-wise alignment suitable for supervised training and evaluation.

view (FoV) and equivalent focal length, and (iii) the maximum fabricable optical thickness. Within these constraints, we derived a single planoconvex micro-lens profile with 20 μm of vertex height and 860 μm radius of curvature that maximizes the effective numerical aperture (NA).

The unit profile was randomly tiled over a 3.5 mm $\times$ 3.5 mm area to form the phase mask. A minimum inter-lens spacing $d_{\min}$ was enforced during placement to preserve fill factor and suppress degradation of the effective NA due to mutual overlap and edge clipping.

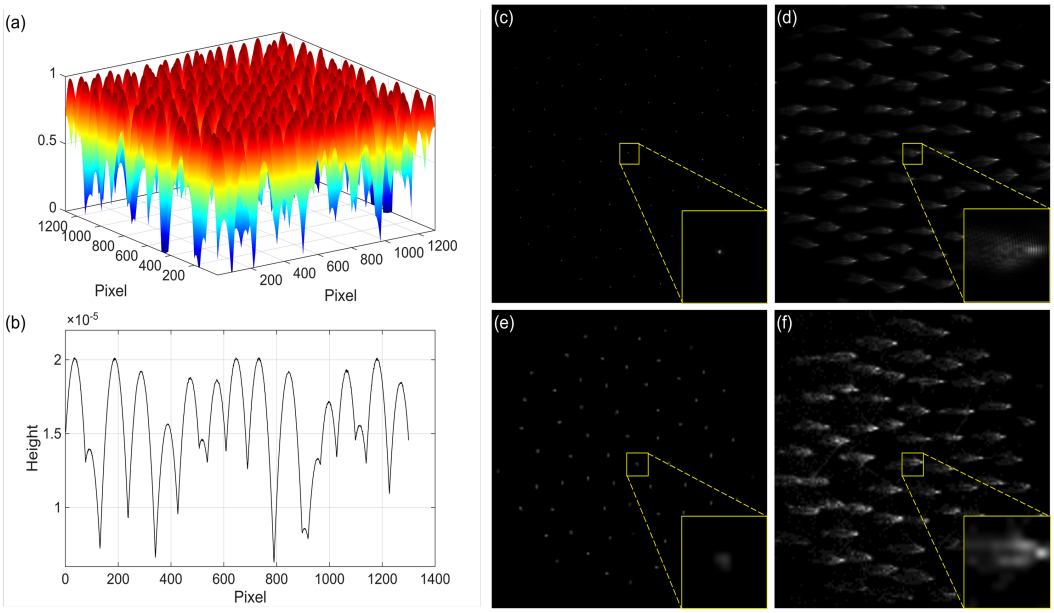

Figure 9: (a) The designed phase mask pattern optimized for high-resolution lensless imaging. (b) A representative line profile. (c) Simulated on-axis PSF (d) simulated off-axis PSF at $40°$, showing the effect of shift variance across the field. (e) captured on-axis PSF (f) captured off-axis PSF showing the effect of shift variance across the field.

FABRICATION

The mask was fabricated as a multi-level phase element via grayscale lithography (Anonymous, 2023). The continuous height map of the planoconvex profile was converted to grayscale dose, enabling a single exposure–development process. Post-fabrication inspection (surface profiler/microscopy) verified profile integrity and absence of large-area defects.

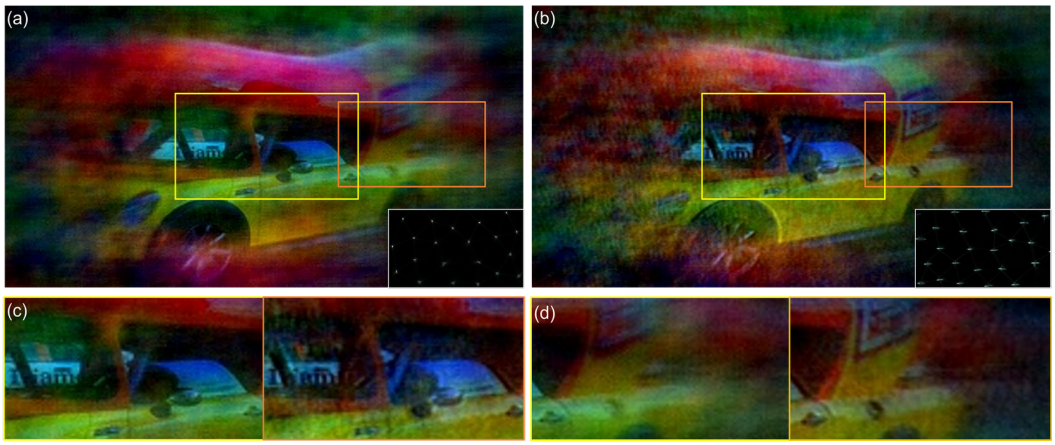

Figure 10: (a) Reconstruction using the center PSF along with the corresponding center PSF. (b) Reconstruction using an off-axis PSF along with the corresponding off-axis PSF. (c) Comparison of sharpness at the image center between (a) and (b). (d) Comparison at the image periphery, demonstrating differences caused by field-dependent PSFs.

Camera Assembly

The fabricated mask was aligned and bonded in front of the RPI3 image sensor to form a lensless camera. A mechanical aperture of 2 mm diameter was applied directly at the mask plane to define the active pupil and to mitigate stray light and edge effects during imaging.

PSF Measurement and Optical Performance

We characterized the system by measuring the point spread function (PSF) on- and off-axis. A comparison between the designed mask, simulated PSFs, and experimentally captured PSFs is shown in Figure 9, highlighting the agreement between design and physical performance.

- **On-axis.** The measured PSF closely matched the designed PSF. Cross-correlation analysis with the design yielded high similarity, and the measured full width at half maximum (FWHM) was small, consistent with the high effective NA. These results provide an empirical bound on the on-axis optical resolution.
- **Off-axis.** At larger field angles, aberrations became pronounced as expected under high-NA operation. The PSF exhibited asymmetric tails consistent with the *coma aberration* from the planoconvex element. Consequently, the system exhibits *shift-variant* imaging behavior: high optical performance on-axis, with aberration-limited quality off-axis.

Implications

Maximizing NA under a limited thickness budget (20 μm) was effective for on-axis resolution but increases sensitivity to off-axis aberrations and FoV non-uniformity. Hardware routes to mitigate this include aspheric refinements, multi-layer (stacked) phase designs, and orientation-aware cell geometries; software routes include deconvolution with a field-dependent PSF or learned reconstructions that explicitly model shift variance. As illustrated in Figure 10, deconvolving with PSFs drawn from different field regions produces noticeable differences in focal sharpness (and associated artifacts), directly evidencing the field dependence.

A.5 Comparison of Deconvolution and ISO

In Figure 11, we compare Wiener deconvolution with our learned inverse system operator (ISO). In DiffuserCam dataset, deconvolution can already recover much of the off-axis information. In contrast, for the SV lensless dataset, deconvolution is less effective due to shift variance from the

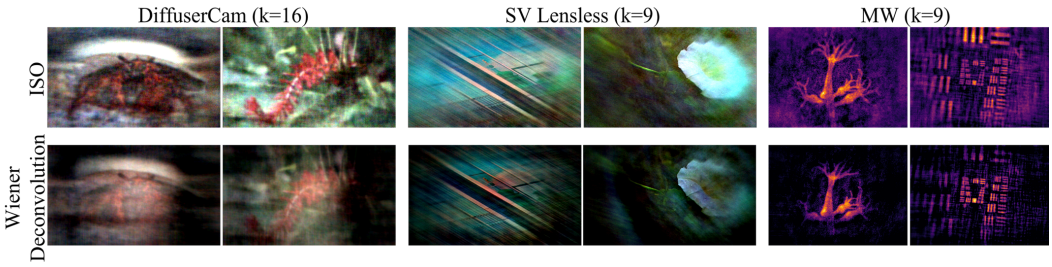

Figure 11: Deconvolution result comparison via inverse system operator proposed in this paper and Wiener deconvolution

Table 2: Quantitative comparison (PSNR ↑ / SSIM ↑ / LPIPS ↓) on Gaussian deblur task with different noise levels σ.

Method	$\sigma = 5$			$\sigma = 10$			$\sigma = 15$		
	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑
RCAN	25.303	0.220	0.810	22.623	0.330	0.750	21.280	0.355	0.711
NAF	**25.625**	**0.205**	**0.825**	**22.810**	**0.315**	**0.766**	21.535	0.345	0.715
Proposed	25.100	0.230	0.800	22.732	0.337	0.741	**21.654**	**0.340**	**.721**

wide FoV design and angular response effects during acquisition, which cause peripheral light loss. The proposed ISO module compensates for these limitations and yields improved reconstructions. On the MW dataset, the ISO recovers a wider range of simulation features and USAF target patterns, demonstrating the benefit of incorporating system-aware operators. Beyond reconstruction, the proposed inverse system operator can function as a standalone pretrained direct inverse mapping function, enabling fast network-based inference.

A.6 Deconvolution Results

A.7 Learned PSFs without Calibration

Figure 12 visualizes the learned 3×3 PSF field from MW dataset. Near the optical center, kernels are compact and approximately isotropic, whereas off-axis locations exhibit increased spread, slight centroid shifts, and mild anisotropy—patterns commonly observed with diffusers and wide-aperture optics. This spatial trend correlates with the improvements seen on real data: FSO reproduces location-dependent blurs using the learned PSFs, and ISO inverts them with data-driven regularization, leading to sharper reconstructions with fewer boundary artifacts.

Notably, the learned PSFs remain normalized and vary smoothly across neighbors, reflecting physically plausible optics. Because the PSF field is shared across scales and injected into every IFIB, the network preserves forward–inverse consistency throughout the hierarchy. Qualitatively, these PSFs agree with expected diffuser patterns and reveal off-axis blur variations that standard shift-invariant models fail to capture, explaining IFIN's robustness under strong shift variance.

A.8 Gaussian Deblur Simulation.

To assess robustness beyond lensless settings, we synthesize a dataset with strong, non-lensless optical blur. Clean natural images are convolved with Gaussian PSFs with $\sigma \in \{5, 10, 15\}$ to emulate heavy defocus. The dataset comprises $24{,}000$ training images and $1{,}000$ test images at 256×256 resolution, isolating the network's ability to handle large kernel support in a conventional deblurring task. See Table 2 for quantitative results. Compared with two strong baselines—RCAN and NAFNet—our method achieves competitive performance at mild blur and becomes increasingly advantageous as blur severity grows. In more realistic scenarios where the blur kernels are better conditioned for deconvolution, as in our main experiments, the benefits of proposed method are expected to be even greater.

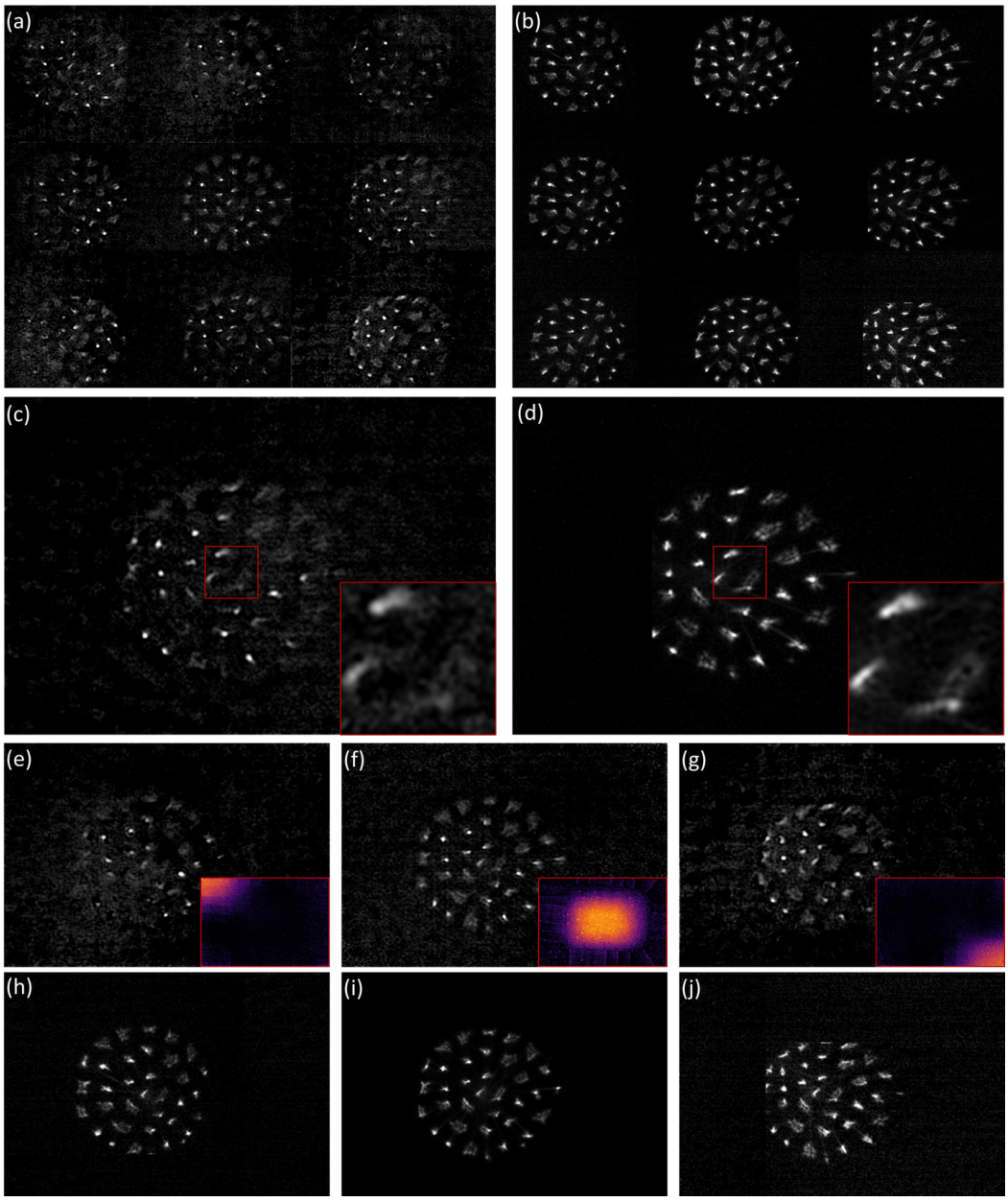

Figure 12: Comparison of learned and calibrated PSFs in MultiWienerNet Dataset (a) PSFs estimated through training. (b) Calibrated PSFs. (c) Estimated PSF at $r = 6$ near the right-center position. (d) Calibrated PSF at the corresponding location. (e–g) Estimated PSFs at different positions according to the learned ROI weights. (h–j) Calibrated PSFs corresponding to the same indices as in (e–g).

A.9 Expansion of System Variance

Depending on the optical configuration, the same PSF field can be re-indexed along the axis.

Concretely, we write

$$h_r[\Delta i, \Delta j;\, q], \quad r = 1, \ldots, s^2, \;\; q \in \mathcal{Q},$$

where r indexes lateral regions (field dependence) and q indexes the nonlateral axis (e.g., depth z, wavelength λ, or time t). For scenarios where the PSF varies both laterally and axially, we adopt a

5-D depth–space-variant PSF:

$$h_{i,j}[\Delta k, \Delta \ell;\, z],$$

The corresponding forward model is

$$y[i,j] \;=\; \sum_{z=1}^{N_z} \sum_{k,\ell} h_{i,j}[k,\ell;\, z]\; x[k,\ell,z] \;+\; \eta[i,j], \tag{18}$$

where $x[\cdot,\cdot,z]$ denotes the scene slice at depth z. for purely field-dependent blur we set $|\mathcal{Q}|{=}1$ and recover the 2-D case. This re-indexing keeps the forward/inverse operators unchanged in form while allowing IFIN to adapt the PSF dimension to the underlying system; in this work we focus on the 2-D shift-variant setting.

A.10 Modeling Shift Variance in FSO

We also considered a fully shift-variant formulation of the forward operator, analogous to ISO, by decomposing $\hat{x}$ into overlapping tiles, padding each tile to the local PSF support, convolving locally, and reassembling via normalized overlap–add:

$$\tilde{y} \;=\; \sum_{r=1}^{m} S_r\Big(\big(R_r\hat{x}\big) * h_r\Big), \tag{19}$$

where R_r extracts the r-th tile and $S_r = R_r^{\top}$ denotes overlap–add with normalization by the local coverage count to avoid seams. In practice, however, this design significantly increases computational overhead despite efforts to optimize tiling. Moreover, the forward operator in IFIN primarily serves to preserve measurement-domain properties rather than to synthesize high-fidelity outputs. Providing it in a simplified, shift-invariant (averaged) form reduces model mismatch while retaining the necessary measurement consistency signal. For these reasons, we adopt the shift-invariant forward operator in our main design. We anticipate that more precise yet simplified variants of the forward operator can be integrated when available, further improving fidelity without incurring significant overhead. A detailed flow of the system operators, including the shift-variant FSO, is provided in Figure 13.

A.11 Ablation Study

Effect of FSO and ISO & Bidirectional Feature Exchange. We assess the contribution of explicitly modeling the forward and inverse processes (FSO/ISO) and the role of bidirectional exchange between them (see Table 3).

FSO/ISO as identity. We replace both FSO and ISO with identity mappings and allow only latent mixing between the two streams. This control demonstrates that *simple feature mixing is not sufficient*: improvements are limited compared to full IFIB, highlighting that enforcing forward and inverse operators *within the feature space* is crucial for propagating measurement-domain cues and stabilizing high-frequency reconstruction.

w/o ISO. We remove the inverse operators and retain only a forward-guided pathway. Without ISO (i.e., deconvolution followed by a denoiser backbone), information flows primarily from the measurement domain. This setting reduces feature sharpening and attenuates high-frequency components, resulting in softer textures and lower PSNR/SSIM.

w/o FSO. We remove the forward operators and retain only inverse guidance, where the measurement is injected only at the first stage of the network. Without FSO to enforce explicit physics in the feature space, the model shows a noticeable drop in fidelity and less stable convergence, although it still benefits from the learned inverse pathway.

Effect of the Number of PSFs We vary the number of learnable PSFs $k = s^2$. Very small k (e.g., $s{=}2$) underfits spatial variability, while excessively large k increases computation without proportional gains. In particular, because the computational cost of the ISO branch scales linearly with k, selecting k becomes especially important when the PSF kernels have large spatial support. We find a sweet spot (e.g., $s{\in}\{3,4\}$ for our image sizes) that balances fidelity and efficiency; details are provided in Table 4.

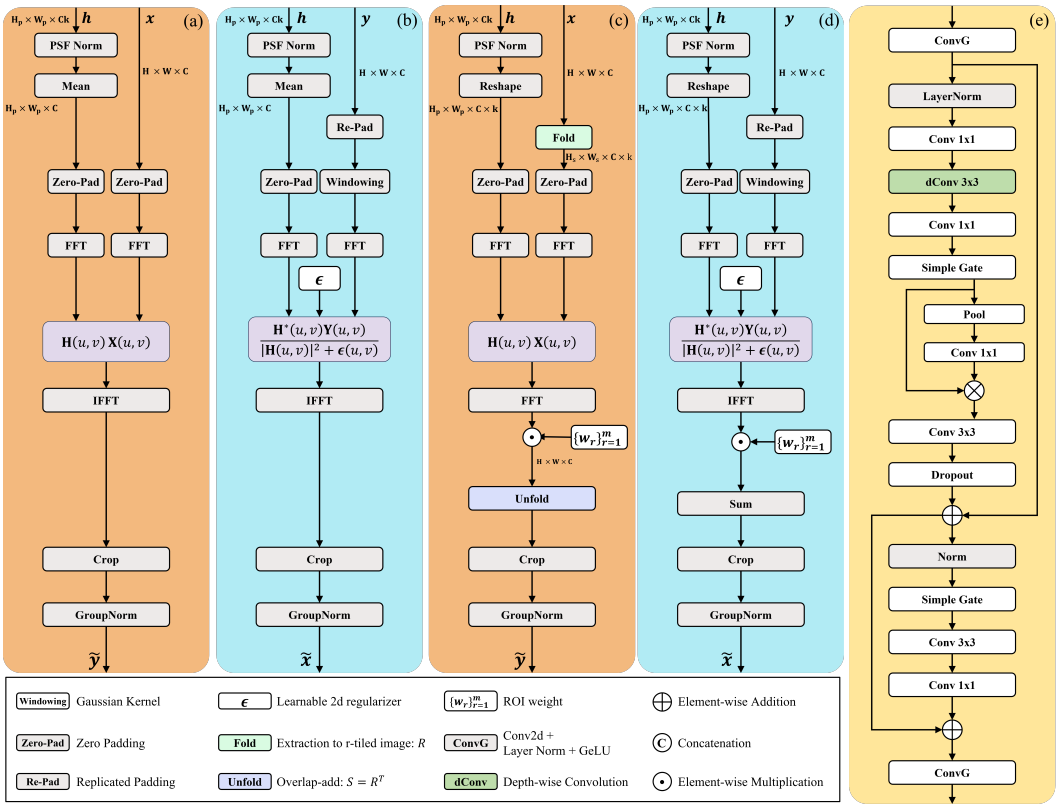

Figure 13: **Detailed flow of system operators and refinement block.** (a) Basic FSO. (b) Basic ISO. (c) Shift-variant FSO. (d) Shift-variant ISO. (e) Refinement block.

Table 3: Comparisons of identity mapping, w/o ISO, and w/o FSO variants show that forward–inverse modeling with bidirectional exchange yields the best fidelity and stability.

Method	DiffuserCam		
	PSNR ↑	LPIPS ↓	SSIM ↑
ISO / FSO as Identity	24.674	0.255	0.800
w/o ISO	27.123	0.223	0.833
w/o FSO	28.711	0.185	0.882
proposed	**29.730**	**0.180**	**0.888**

Effect of the Initial Guess of PSFs We compare initializing the PSF with the calibrated on-axis measurement versus random noise. Using the calibrated PSF provides a consistent PSNR gain and accelerates convergence, likely because a coarse kernel estimate is available from the start. Importantly, however, calibration is not strictly required: even without access to an on-axis PSF, as long as the forward system is well defined and training pairs are available, the network can learn to estimate effective kernels in a fully data-driven blind deconvolution setting. In such cases, the performance drop is modest, and the network still converges to meaningful PSFs. While initialization with a calibrated PSF yields faster and more stable training, the ability to recover kernels without calibration underscores the robustness and practical utility of the approach.

Table 4: quantitative comparison (PSNR / SSIM) on four benchmark datasets

Method	**DiffuserCam**		
	PSNR ↑	LPIPS ↓	SSIM ↑
k=1 (SI)	28.840	0.190	0.864
k=4	29.112	0.186	0.880
k=9	29.484	0.182	0.885
k=16	29.730	0.180	0.888

Table 5: quantitative comparison (psnr / ssim) on four benchmark datasets

Method	**DiffuserCam**		
	PSNR ↑	LPIPS ↓	SSIM ↑
initialize as random	28.984	0.189	0.872
initialize as center PSF	29.730	0.180	0.888

A.12 Discussion and Limitation

Key Takeaways. Across the three benchmarks (DiffuserCam, SV Lensless, and MultiWienerNet), IFIN sets a new state of the art, surpassing the strongest learning baselines by +1.50 dB, +0.92 dB, and +3.65 dB PSNR, respectively—averaging +2.02 dB PSNR, -0.033 LPIPS, and $+0.040$ SSIM. Gains are especially clear near the field periphery and on MW, where large PSFs degrade purely CNN-based inversion. We attribute these improvements to (i) integrated forward–inverse coupling, which enforces measurement-domain consistency, and (ii) learnable shift-variant operators that utilize a learnable PSF field to adapt to system mismatch.

Accuracy vs. Computational Cost. Spatially varying operators improve fidelity but reduce parameter sharing, increasing memory usage and runtime relative to shift-invariant (SI) models. IFIN mitigates this by cropping PSFs to regions with signal and multi-scale processing to avoid unnecessary computation.

On inputs of size $H\times W{=}270\times480$ and a 3-scale IFIN, throughput decreases as the number of learnable PSFs k grows: ≈ 0.87 MP/s at $k{=}1$ (148.1 ms), 0.76 MP/s at $k{=}4$ (171.6 ms), 0.58 MP/s at $k{=}9$ (223.7 ms), and 0.46 MP/s at $k{=}16$ (282.3 ms) on a single NVIDIA A6000 GPU. Peak memory usage scales similarly, from 0.66 GB at $k{=}1$, 1.95 GB at $k{=}4$, 4.13 GB at $k{=}9$, to 7.26 GB at $k{=}16$ (allocated, with total reserved up to 11.3 GB).

Generality Beyond a Single Modality. The integrated forward–inverse design is agnostic to a specific hardware stack. Any system that admits a (possibly varying) convolutional forward model—including defocus/motion blur in photography, field-dependent aberrations in microscopy/telescopy, or turbulence in astronomy—can benefit from IFIN's combined forward–inverse modeling with flexible system adaptation.

Potential Extensions. Promising directions include depth- and wavelength-dependent operators for volumetric and hyperspectral imaging, low-rank/separable PSF factorization for speed, and coordinate-conditioned or deformable operator parameterizations that retain variance with fewer weights. Hardware–algorithm co-design (e.g., masks/apertures yielding sparser or more localized PSFs) could further reduce overlap and ringing.