

Integrated Forward–Inverse Network for Lensless Image Reconstruction

Donggeon Bae¹, Jaewoo Jung^{2,3}, Yong Guk Kang², Kyung Chul Lee⁴,
Taeyoung Kim³, Jongho Kim¹, Sangjun Byun¹, Joonsik Park³, and
Seung Ah Lee^{1,2}*

¹ Department of Mechanical Engineering, Seoul National University, Republic of Korea

² School of Mechanical and Aerospace Engineering/SNU-IAMD, Seoul National University, Republic of Korea

³ Department of Electrical and Electronic Engineering, Yonsei University, Republic of Korea

⁴ Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI, USA

{donggeonbae, seungahlee}@snu.ac.kr

Abstract. Lensless imaging enables compact and versatile computational cameras by replacing bulky optics with thin coded elements. However, reconstruction from the resulting measurements is challenging: large-footprint point-spread functions (PSFs) produce highly multiplexed observations, making inversion severely ill-conditioned and sensitive to calibration errors and model mismatch. While deep learning approaches, including hybrid models that incorporate physics priors, have shown promise, explicitly maintaining data fidelity throughout the network hierarchy remains difficult. Here, we propose the Integrated Forward–Inverse Network (IFIN), a physics-guided architecture that interleaves differentiable forward projections with learnable inverse updates at every scale, enabling complementary cues to be exploited jointly in the measurement and image domains. This bidirectional coupling supports progressive, physics-consistent refinement and permits system-constrained PSF kernel adaptation under model uncertainty. On challenging lensless benchmarks, including a newly introduced dataset, IFIN achieves state-of-the-art reconstruction quality. We further observe competitive performance on Gaussian deblurring and simulated inline holography reconstruction, suggesting that the same interleaving principle can extend beyond lensless cameras.

Keywords: Computational Imaging · Lensless Imaging · Model-based Deep Learning · Inverse Problems

1 Introduction

Modern optical imaging systems, ranging from compact lensless cameras with coded apertures to advanced microscopes with engineered point-spread functions

* Corresponding author.

(PSFs), are increasingly designed with complex forward models. A growing class of such systems operates with PSFs that are intentionally or unavoidably broadened by designed optical coding. In this regime, each measurement mixes scene information over a wide spatial extent, making the inverse problem severely ill-conditioned. While such designs unlock diverse imaging capabilities that transcend the limits of conventional optics [1, 2, 5, 40, 41], they also introduce substantial challenges for reconstruction. In practice, the effective PSFs can vary across both the field of view and channels, violating the stationarity assumptions underlying standard inverse pipelines [13, 42, 49]. Hardware imperfections and residual modulations further deteriorate this model mismatch, making accurate and robust reconstruction a central challenge as optical platforms continue to shrink and diversify.

A wide spectrum of approaches has been explored for image reconstruction in lensless imaging systems. Classical inverse mappings [45] and model-based optimization methods [12, 28, 38] built on well-defined priors offer physically grounded results, but they are often computationally expensive, sensitive to calibration errors, and unreliable under model mismatch. With advances in deep learning, data-driven methods [4, 35] have enabled end-to-end mappings from measurements to target scenes, yet they may not explicitly encode the underlying system physics, which can reduce accuracy and robustness under out-of-distribution conditions; such models may also produce hallucinations. In response, hybrid methods [24, 26, 34, 49] 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.

However, many existing hybrid pipelines incorporate physics in a one-sided manner, either operating primarily in the measurement domain or refining only an already inverted estimate; once an image is reconstructed, the measurement cues become less accessible. Consequently, measurement information may collapse as it passes through the network, and intermediate estimates can become detached from the raw measurements. This becomes especially problematic in practical lensless reconstruction, where the effective forward model varies across the field of view and is only approximately calibrated. In this setting, whether inversion is applied after this collapse or refinement follows an inversion without measurement-domain feedback, mismatch-induced errors can manifest as spatially structured artifacts and persist through later stages.

A key opportunity in this setting is to leverage physics priors and learned priors jointly throughout reconstruction, rather than confining physical consistency checks to a single stage. In optical systems with broad PSFs and long-range mixing, preserving measurement-domain cues while refining image-domain representations provides complementary information for recovering fine details. To this end, we introduce **IFIN**, a bidirectional reconstruction framework that interleaves differentiable forward projections with learnable inverse updates within an encoder–decoder hierarchy, enabling stable, physics-consistent reconstruction under large-footprint PSFs and shift-variant degradations. IFIN further learns a

shift-variant PSF field end-to-end, improving robustness to calibration mismatch and enabling blind recovery when PSFs are inaccurate or unavailable.

Compared to the prior state of the art, IFIN improves PSNR on the three lensless benchmarks by +1.63 dB (DiffuserCam [34]), +0.65 dB (**WiderCam**, our newly introduced lensless benchmark), and +2.58 dB (MultiWienerNet [49]). Beyond lensless imaging, IFIN remains competitive on simulated Gaussian deblurring and achieves strong gains on inline holography reconstruction, demonstrating that the proposed bidirectional forward-inverse integration generalizes beyond a single modality and can be applied to a broader class of inverse problems.

Our main contributions are as follows:

1. We propose IFIN, a reconstruction framework that embeds *bidirectional* forward-inverse guidance at *every* encoder-decoder scale, repeatedly exchanging measurement- and image-domain cues for physics- and data-driven refinement under long-range mixing.
2. Within this framework, IFIN jointly learns a shift-variant PSF field, shared by both operators across scales, for robustness to calibration mismatch and blind recovery.
3. IFIN achieves state-of-the-art performance on three lensless benchmarks and further validates the proposed approach on simulated Gaussian deblurring and inline holography.

2 Related Work

2.1 Lensless Imaging

Lensless cameras replace conventional lenses with thin optical elements such as coded apertures [3], transmissive diffusers [1], and engineered phase masks [10, 25]. As a result, diffuser- or mask-induced PSFs are large and highly structured, often encoding wide spatial neighborhoods onto the sensor, up to the entire scene. This encoding eliminates the need for bulky optics but necessitates computational reconstruction to recover interpretable images from the raw measurements.

Beyond simple image recovery, lensless systems support diverse modalities—including depth [1, 6], hyperspectral [33, 40], polarization [5], ultrafast video via rolling-shutter coding [2], and privacy-preserving imaging [21, 41]—making them appealing for embedded vision under strict size and cost constraints [20, 23, 46].

Image reconstruction becomes particularly challenging when the optical system produces highly multiplexed measurements, often modeled as 2D or 3D convolutions. 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 tasks such as imaging through scattering media [50],

non-line-of-sight imaging [19], coherent diffractive imaging [30], and microscopy with engineered PSFs [36].

We begin with a baseline shift-invariant model, where the system response is identical across locations and the measurement is a 2D convolution between the scene and a static PSF:

$$y[i, j] = \sum_{a, b} h[a, b] x[i - a, j - b] + \eta[i, j], \quad (1)$$

where $x, y \in \mathbb{R}^{H \times W}$ denote the scene irradiance and the captured measurement, h is a static PSF of the system, and η models additive noise.

In practice, most imaging systems are not truly shift-invariant, even though they are often modeled as convolutions. Off-axis aberrations, depth-dependent propagation, field-dependent magnification, vignetting or pupil clipping, and sensor truncation all make the effective system response depend on spatial location [1, 11, 42]. This is especially pronounced for phase or coded masks with high effective numerical aperture: resolution improves on-axis, but aberration-induced shift variance grows with field angle. As a result, the location-dependent PSF $h_{i,j}$ at pixel position $[i, j]$ 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:

$$y[i, j] = \sum_{a, b} h_{i,j}[a, b] x[i - a, j - b] + \eta[i, j], \quad (2)$$

The forward model is no longer a 2D convolution but a large, spatially varying operator, raising the computational and memory cost of precise inversion; sensor cropping and noise further increase ill-posedness and calibration sensitivity.

2.2 Image Restoration

Given such forward models, image recovery in lensless cameras is carried out through computational inversion, often posed as deconvolution, closely related to reconstruction in conventional cameras where strong degradation can arise from aberrations, motion, or turbulence. Classical estimators such as Wiener filtering [45] and Richardson–Lucy iterations [28, 38] provide long-standing baselines and are attractive when the forward model is accurate, but their practical performance is often limited by noise amplification and PSF misspecification. To handle non-ideal forward models that include truncation/cropping and non-smooth priors (e.g., total variation, TV), iterative optimization frameworks such as ADMM [12] are widely used to decouple data fidelity from regularization and enable tractable solvers [1].

In parallel, deep networks, including CNN-based restorers [4, 15, 39, 52] and ViT architectures [18, 35], learn direct mappings from measurements to images and have become strong empirical baselines for large-scale restoration. Motivated by their complementary strengths, a growing line of work integrates explicit

forward-model structure into learning-based pipelines. Examples include reconstruction via unrolled iterations with learned denoisers [9, 24, 34, 37], measurement-consistency constraints for unsupervised training [32, 43, 44], or feed-forward hybrids that combine a physical inversion stage with learned refinement [22, 49]. Related approaches embed deconvolution within multi-scale feature hierarchies to improve fidelity and robustness [7, 17, 26]. The above methods differ primarily in how they use the measurements: many enforce physics through a fixed forward-model constraint or through a single inversion step, after which the measurement-domain information is no longer explicitly carried through the hierarchy. Since inversion difficulty depends strongly on the PSF support and the conditioning of the forward operator, we next formalize the forward model and summarize three fundamental failure modes that emerge as the PSF support grows and model mismatch increases. This motivates architectures that preserve and update measurement- and image-domain representations in tandem.

3 Key Challenges

We consider a linear forward model $y_\gamma = H_\gamma x + \eta$, where $x \in \mathbb{R}^N$ is the scene, $y_\gamma \in \mathbb{R}^M$ the measurement, $H_\gamma \in \mathbb{R}^{M \times N}$ a forward operator parameterized by complexity γ , and η sensor noise. As γ grows, the PSF widens and mixes information across increasingly distant pixels, creating three failure modes for neural restoration that motivate the design of IFIN.

Locality Mismatch Under Degradation. A wide-support H_γ aggregates long-range contributions into each measurement, whereas CNNs and ViTs process y_γ with finite receptive fields or windowed self-attention [27, 29]. When the kernel support exceeds a layer’s effective field, out-of-field contributions behave as structured noise, amplifying the difficulty posed by H_γ through the mismatch between its large footprint and the network’s locality.

Loss from Model-Based Inversion. A second source of degradation arises from how the inverse problem is solved. When H_γ is ill-conditioned, the consistent set $\mathcal{S}(y_\gamma) = \{x \in \mathbb{R}^N : H_\gamma x \approx y_\gamma\}$ is high-dimensional, so weakly constrained high-frequency directions are nearly unobservable and classical or learned inverses resolve this by selecting a single smooth estimate $\hat{x} = G(y_\gamma)$ [8, 14]. One-sided physics-NN pipelines that feed only $\hat{x}$ into a network [9, 22, 24, 34, 37, 49] discard the measurement-domain residuals that show how $\hat{x}$ fails to explain y_γ , leaving fine detail to learned priors rather than data consistency.

Representation Bottlenecks. Even with inversion embedded in dimension-reducing encoders $z_\gamma = f_\theta(y_\gamma)$ [7, 17, 26], non-invertibility gives $I(x; z_\gamma) \leq I(x; y_\gamma)$ [16], and as γ grows the increasingly ill-conditioned, low-rank H_γ pushes more fine-detail components below the encoder’s effective thresholds. Inserting

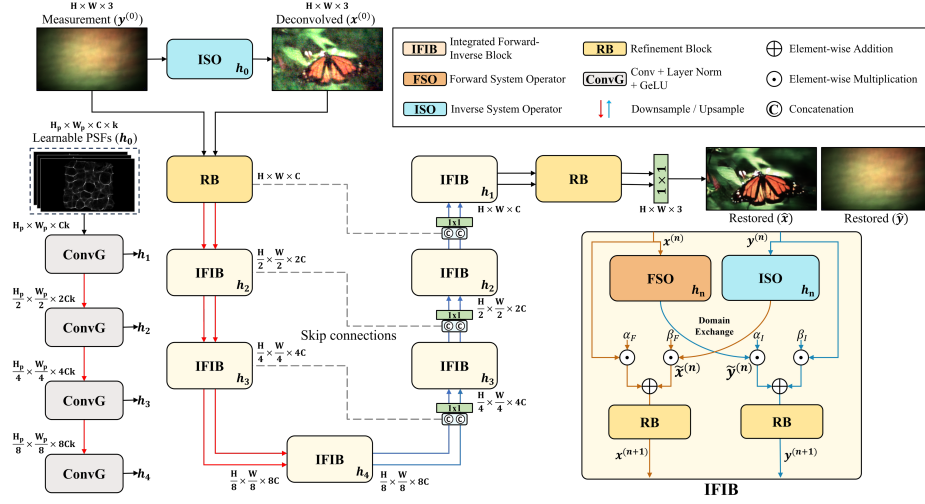


Fig. 1: Overall architecture of IFIN. The network follows an encoder–decoder structure, where Integrated Forward–Inverse Blocks (IFIBs) are inserted at each scale to jointly apply the Forward System Operator (FSO) and Inverse System Operator (ISO). A shared learnable PSF field guides both operators, ensuring forward–inverse consistency across scales. Notation: n indexes the per-scale PSF embedding h_n (one per scale, shared across the hierarchy); (n) denotes the stage- n representations $(x^{(n)}, y^{(n)})$ updated across the hierarchy.

inversion alone is therefore insufficient: unless the architecture explicitly preserves and propagates measurement information across layers, it remains vulnerable to the combined effects of ill-conditioned low-rank forward models and representation bottlenecks.

4 Method

The proposed IFIN adopts an encoder–decoder backbone equipped with Integrated Forward–Inverse Blocks (IFIBs) at every scale (Fig. 1). The encoder progressively downsamples the input measurement and coarse estimation to capture coarse-scale coupling and long-range interactions, while the decoder upsamples features back to the native resolution to recover fine details. At each resolution, an IFIB couples a Forward System Operator (FSO), which maps image-domain features to the measurement domain, with an Inverse System Operator (ISO), which restores image-domain features from the measurement. Across scales, both operators condition on and jointly refine a learnable PSF field, enforcing forward–inverse consistency while propagating physically meaningful residual signals throughout the network.

At the input stage, a coarse estimation is obtained by applying the ISO to the measurement. The pair (*measurement*, *coarse estimation*) is then propagated as

two coupled streams through the encoder-decoder hierarchy. Within each IFIB, the FSO and ISO exchange features bidirectionally, jointly enforcing consistency across the measurement and image domains.

4.1 Learnable PSF Field

IFIN incorporates a learnable PSF representation that provides explicit system awareness to both the FSO and the ISO. The PSF field is parameterized as $k=s^2$ kernels covering local regions of the image. When $s=1$, the PSF field reduces to a single global kernel. Kernels can be initialized from calibrated measurements, a single reference PSF, or random patterns, and are jointly optimized end-to-end within the network. In our experiments we use $k \in \{1, 4, 9, 16\}$ for DiffuserCam and $k=9$ for WiderCam and MultiWienerNet (MWNet), with cropped PSF supports of 270×270 , 135×135 , and 320×224 for the three lensless benchmarks (and 101×101 for Gaussian deblurring).

A compact PSF encoder maps the field to multi-scale embeddings $\{h_n\}$, where each h_n is shared between the encoder and decoder IFIBs at scale n and injected into both the FSO and the ISO, maintaining physical consistency across the hierarchy. Because the same PSF field must explain both how features generate measurements (FSO) and how measurements invert to sharp images (ISO), the PSFs are constrained by complementary supervision in both domains, which improves identifiability, discourages degenerate kernels, and aids blind PSF estimation.

We normalize the PSF to obtain unit DC gain in both the FSO and the ISO. This stabilizes the physics operators and prevents scale drift. We further impose a weak non-negativity regularizer on the PSF by penalizing negative values during training, guiding the learned kernels toward physically plausible solutions.

4.2 Integrated Forward-Inverse Block (IFIB)

The IFIB is the fundamental unit of IFIN, coupling forward and inverse imaging at each scale through two parallel operators, FSO and ISO (Fig. 2), both tied to the system’s physics. They can be configured as shift-invariant when degradations are approximately uniform, or spatially varying for more complex degradations, letting IFIN balance efficiency and fidelity.

Forward System Operator (FSO). The FSO simulates how the current image-domain estimate x would be formed by the imaging system, producing a forward projection in the measurement domain. By default, we use 2D linear convolution with zero padding via a single PSF h :

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

The projection $\tilde{y}$ is a measurement-domain proxy induced by the current image representation. Inside each IFIB it augments the measurement-domain

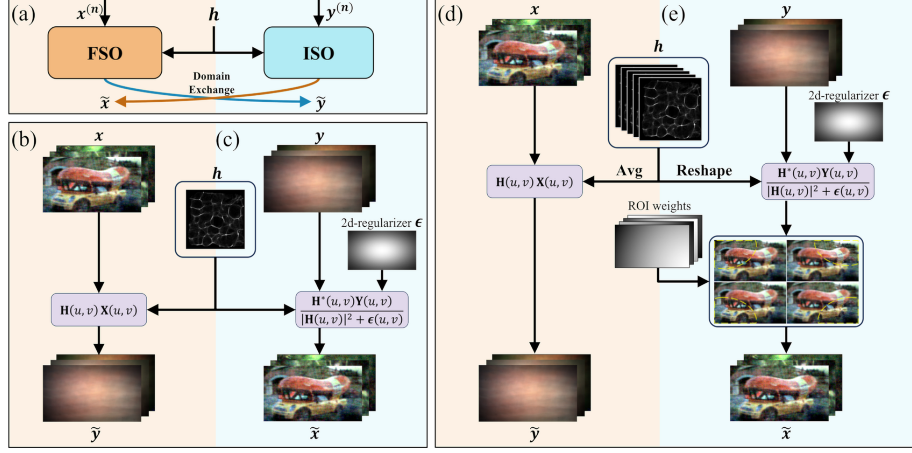


Fig. 2: (a) Schematic of the forward-inverse pairing in the IFIB. (b,c) **Single-PSF setting:** FSO uses 2D convolution; ISO uses Wiener-like deconvolution. (d,e) **PSF-field setting:** FSO uses a single representative (averaged) PSF from the PSF field; ISO applies region-wise deconvolution blended by learnable region-of-interest (ROI) maps.

features to improve their consistency with the forward model and to better support the subsequent inverse update. For efficiency under large-footprint PSFs, we compute it in the frequency domain with zero padding and crop back to the native resolution. When the PSF field provides multiple kernels, the FSO uses a single representative (averaged) kernel for this projection. Because its role is to supply a stable measurement-consistency cue rather than to synthesize a high-fidelity measurement, an averaged kernel reduces model mismatch and avoids the cost of region-wise forward projection. In contrast, the ISO carries out the locally sensitive deconvolution and therefore applies region-wise inversion with ROI blending (below) to capture field-dependent degradation.

Inverse System Operator (ISO). The ISO restores a sharp estimate from the degraded measurement via Wiener-like deconvolution with a learnable frequency-dependent regularizer. For PSF h , letting $Y(u, v) = \mathcal{F}\{W \cdot P_{rp}y\}(u, v)$ 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 [22], we compute:

$$\hat{X}(u, v) = \frac{H^*(u, v)}{|H(u, v)|^2 + \epsilon(u, v)} Y(u, v), \quad \epsilon(u, v) \geq 0, \quad (4)$$

and set $\hat{x} = \mathcal{F}^{-1}\{\hat{X}\}$. Here, $\epsilon(u, v)$ is a learnable 2D parameterization refined during training, with non-negativity enforced by a ReLU. After the inverse Fourier transform, we crop the result back to the native resolution, matching the forward

projection size. Within each IFIB, this reconstructed estimate is forwarded as an augmentation signal to the subsequent image-stream refinement.

When the PSF field provides multiple kernels, we can apply the same inverse Eq. (4) region-wise using PSFs $\{h_r\}_{r=1}^k$ and regularizers $\{\epsilon_r\}_{r=1}^k$, and blend the reconstructions with normalized ROI weights to capture local variability:

$$\hat{x}[i, j] = \sum_{r=1}^k w_r[i, j] \mathcal{F}^{-1}\{\hat{X}_r\}[i, j], \quad (5)$$

where $\hat{X}_r$ denotes the result of Eq. (4) computed with the region-wise pair (h_r, ϵ_r) , and $\{w_r\}_{r=1}^k$ are k learnable ROI maps. We initialize the ROI maps from Gaussian kernels $\{g_r\}_{r=1}^k$ centered at p_r , where $k=s^2$ and p_r are the centers of an $s \times s$ grid partitioning the input measurement:

$$g_r[i, j] = \exp\left(-\frac{\|(i, j) - p_r\|_2^2}{2\sigma_r^2}\right), \quad w_r[i, j] = \frac{g_r[i, j]}{\sum_{q=1}^k g_q[i, j]}, \quad (6)$$

$$\sum_{r=1}^k w_r[i, j] = 1 \quad \forall (i, j), \quad (7)$$

where σ_r is the width of the r -th initialization Gaussian.

Integrated Forward-Inverse. The hallmark of the IFIB is the bidirectional exchange between FSO and ISO. At each stage (n) , the FSO produces a measurement-domain projection $\tilde{y}^{(n)} = \text{FSO}(x^{(n)}; h_n)$ from the current image-domain representation, while the ISO produces an image-domain estimate (a feature map) $\tilde{x}^{(n)} = \text{ISO}(y^{(n)}; h_n)$ from the current measurement-domain representation. These cross-domain outputs are then used as augmentation signals that are fused into the next updates of both streams:

$$y^{(n+1)} = \phi_\theta^y \left(\alpha_F^{(n)} \cdot y^{(n)} + \beta_F^{(n)} \cdot \tilde{y}^{(n)} \right), \quad (8)$$

$$x^{(n+1)} = \phi_\theta^x \left(\alpha_I^{(n)} \cdot x^{(n)} + \beta_I^{(n)} \cdot \tilde{x}^{(n)} \right), \quad (9)$$

where ϕ_θ^y and ϕ_θ^x each consist of a sequence of convolutional layers, and $\alpha_F^{(n)}, \beta_F^{(n)}$ and $\alpha_I^{(n)}, \beta_I^{(n)}$ are learnable per-channel scalar gates for the measurement and image streams at scale n . Thus $\tilde{y}^{(n)}$ updates the measurement representation before the next ISO step, while $\tilde{x}^{(n)}$ drives the image-domain update and conditions the next FSO projection, forming a forward-inverse coupling across the hierarchy. Layer configurations and comprehensive FSO/ISO ablations are in the supplementary material.

Table 1: Quantitative comparison on three benchmarks—**DiffuserCam**, **WiderCam**, and **MWNet dataset**. **ISO (Ours)** denotes the learned initial inverse operator used for the coarse estimation stage of IFIN. We report PSNR $\uparrow$, LPIPS $\downarrow$ [51], and SSIM $\uparrow$ (arrows indicate the preferred direction). Best in bold, second best underlined.

Dataset	DiffuserCam			WiderCam			MultiWienerNet		
Metrics	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$
ADMM [12]	12.252	0.607	0.346	11.843	0.643	0.323	19.189	0.557	0.420
Wiener Deconv. [45]	12.552	0.591	0.384	12.405	0.607	0.369	18.658	0.640	0.302
ISO (Ours)	16.528	0.544	0.404	17.240	0.462	0.444	20.202	0.623	0.380
U-Net [39]	21.230	0.394	0.656	21.890	0.474	0.646	23.859	0.389	0.589
NAFNet [15]	24.830	0.239	0.810	23.857	0.245	0.769	24.657	0.282	0.712
Le-ADMM-U [34]	23.261	0.312	0.765	21.956	0.278	0.748	23.732	0.335	0.702
DeepLIR [37]	25.958	0.260	0.829	20.523	0.339	0.642	22.556	0.379	0.642
MWNet [49]	24.832	0.247	0.810	23.001	0.255	0.766	25.660	0.260	0.728
UPDN [24]	<u>28.228</u>	0.194	0.877	23.920	0.229	0.801	24.364	0.287	0.707
MWDNs [26]	27.298	0.217	0.845	24.525	0.224	0.801	27.436	0.236	0.780
LensNet [7]	27.650	0.201	0.868	24.615	0.219	0.806	27.546	0.221	0.809
MoDL [9]	27.958	<u>0.183</u>	<u>0.878</u>	<u>24.791</u>	<u>0.202</u>	<u>0.810</u>	<u>28.504</u>	<u>0.202</u>	<u>0.831</u>
IFIN (Ours)	29.862	0.174	0.893	25.444	0.201	0.824	31.083	0.175	0.866

5 Results

We evaluate IFIN via end-to-end supervised training on paired scene-measurement data spanning real display-capture and synthetic settings. We report results on three lensless benchmarks—**DiffuserCam** [34], **WiderCam** (ours), and **MultiWienerNet (MWNet)** [49]—which together cover a widely used diffuser dataset, strong shift variance with a large field of view, and multi-scale PSF fields with real experimental validation. We compare against classical baselines [12, 45], data-driven models [15, 39], and representative hybrid approaches [7, 9, 24, 26, 34, 37, 49]. To probe the scope of IFIN beyond lensless imaging, we additionally include synthetic **Gaussian deblurring** and **diffractive reconstruction** benchmarks. All methods use a unified reconstruction objective with standard method-specific recipes; further dataset, implementation, and training details are in the supplementary material.

DiffuserCam. DiffuserCam [34] contains 25,000 paired display-capture measurements acquired with the DiffuserCam prototype [1] (24,000 train / 1,000 test); we use the standard 480×270 measurements obtained by $4\times$ downsampling from 1920×1080 raw sensor frames.

On this benchmark, IFIN improves PSNR, LPIPS, and SSIM over prior methods (Tab. 1) and yields clearer reconstructions with fewer ringing/haze artifacts while better preserving textures and color consistency (Fig. 3). These results are consistent with the design rationale of IFIN, which repeatedly reintroduces measurement-consistency cues during decoding via forward-inverse refinement.

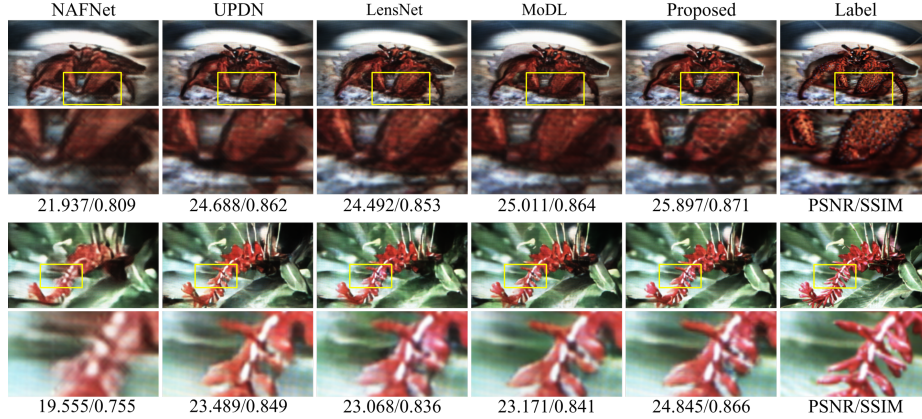


Fig. 3: Visual comparison on DiffuserCam display-capture data. IFIN preserves color fidelity and high-frequency textures while suppressing artifacts. Insets mark zoomed regions and structures.

WiderCam. WiderCam (ours) consists of 25,000 wide field-of-view (FoV, $> 100^\circ$) lensless measurements (24,000 train / 1,000 test), resized to 480×270 from 4608×2592 sensor frames, with affine-aligned supervision. The dataset exhibits strong field dependence: regions near the optical center are typically easier to reconstruct, while peripheral regions are substantially more challenging due to shift variance and non-uniform degradation.

IFIN shows its advantage most clearly in the outer field, recovering sparse peripheral structures more reliably while maintaining sharp central content (Fig. 4), consistent with the quantitative gains in Tab. 1. We further test generalization on in-the-wild captures without ground truth (Fig. 5); IFIN remains stable across diverse scenes and illumination, reducing ringing and color shifts while preserving edge fidelity.

MWNet Dataset. MultiWienerNet (MWNet) [49] provides 22,125 paired 2D samples (17,700 train / 4,425 test) generated using a measured spatially varying PSF field from a mask-based miniscope [48]; in our experiments, we resize images to 320×224 for efficiency. On this benchmark, IFIN improves quantitative metrics (Tab. 1) and produces visibly sharper reconstructions from both synthetic measurements and real USAF captures (Fig. 6). IFIN does not require explicit region-wise PSF initialization: starting from limited calibration (e.g., a single on-axis PSF), it can learn effective position-dependent behavior during reconstruction. Compared to MWNet, which relies on multiple calibrated PSFs (9 in the original setting) to recover off-axis information, IFIN surpasses it by a clear margin while requiring far less calibration. It exceeds MWNet without the learnable PSF field and even without calibrated PSF information, while learning the PSF field end-to-end improves performance further.

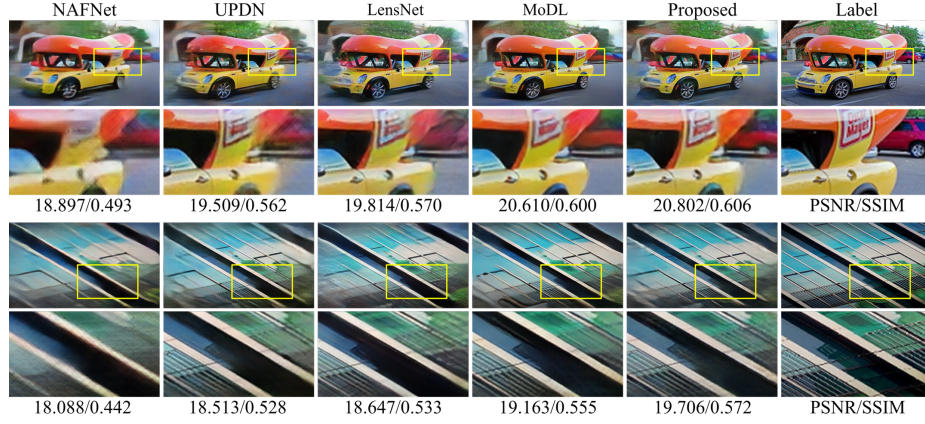


Fig. 4: Comparison on WiderCam dataset. Compared to prior methods, IFIN mitigates field-dependent peripheral blur and geometric distortion, while preserving fine textures and edges across the entire image.

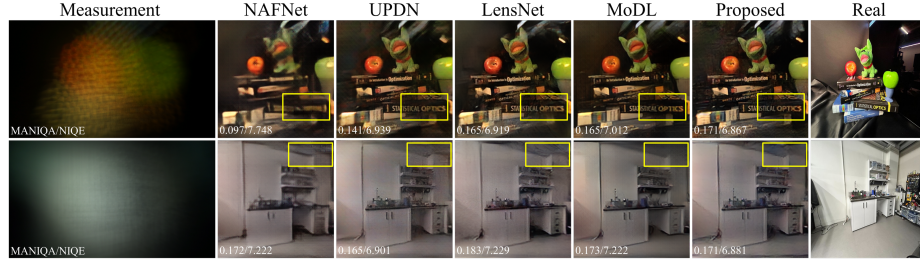


Fig. 5: In-the-wild WiderCam measurements. IFIN generalizes to diverse scenes and lighting, reducing ringing and color shifts while preserving edges and textures. Without ground truth, we additionally report no-reference image quality metrics: MANIQA $\uparrow$ [47] and NIQE $\downarrow$ [31].

Gaussian Deblurring Simulation. In addition to lensless experiments, we evaluate IFIN on synthetic Gaussian deblurring (Tab. 2). We blur 384×384 images with Gaussian kernels of increasing standard deviation σ and compare against strong purely data-driven restorers [15, 52].

For mild blur, NAFNet/RCAN remain highly competitive, yet IFIN achieves comparable performance, indicating that the forward-inverse interleaving prior is not limited to large-footprint degradations. As the blur footprint increases, IFIN becomes increasingly advantageous and yields the best overall performance at $\sigma = 15$ (PSNR/SSIM $\uparrow$, LPIPS $\downarrow$). Overall, these results suggest that incorporating explicit forward and inverse operators helps IFIN degrade more gracefully as the inverse problem becomes more ill-conditioned, while remaining effective in the small-footprint regime.

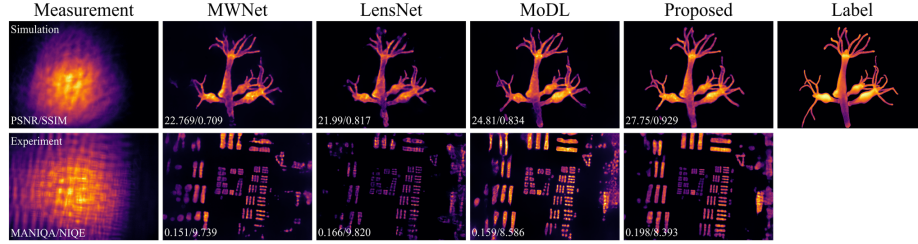


Fig. 6: Comparison on MWNet dataset. The first row shows simulated spatially varying measurements and reconstructions; the second row shows experimental miniscope captures of a USAF target.

Table 2: Gaussian deblurring with standard deviations $\sigma \in \{5, 10, 15\}$. We report PSNR $\uparrow$ /LPIPS $\downarrow$ /SSIM $\uparrow$.

Method	$\sigma = 5$			$\sigma = 10$			$\sigma = 15$		
	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$
RCAN	<u>25.303</u>	<u>0.220</u>	<u>0.810</u>	22.623	<u>0.330</u>	<u>0.750</u>	21.280	0.355	0.711
NAFNet	25.625	0.205	0.825	22.810	0.315	0.766	<u>21.535</u>	<u>0.345</u>	<u>0.715</u>
IFIN (Ours)	25.100	0.230	0.800	<u>22.732</u>	0.337	0.741	21.654	0.340	0.721

Table 3: Inline holography ($z = 30$ mm; $\lambda = 532$ nm).

Method	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$
RCAN	23.764	0.354	0.702
NAFNet	23.224	0.389	0.634
NAFNet+	<u>25.751</u>	<u>0.242</u>	<u>0.817</u>
IFIN (Ours)	28.302	0.166	0.890

Diffraction Imaging Reconstruction. We evaluate whether IFIN transfers to an inverse problem with a fundamentally different forward operator by testing simulated inline holography, a diffractive imaging modality governed by wave propagation. Using the Dogs vs. Cats dataset, we generate holographic measurements from 256×256 images via the angular spectrum method at $z = 30$ mm and $\lambda = 532$ nm, add a mixture of Gaussian and Poisson noise, and reconstruct the amplitude-only transmission object.

To adapt IFIN to inline holography, we keep the interleaving backbone and replace the system operators with modality-matched ones: the FSO is instantiated with forward propagation using the angular spectrum method, and the ISO uses the corresponding back-propagation operator as an inversion module. Since this setting is governed by wave propagation rather than a spatial PSF field, we omit the PSF encoder and do not perform PSF-field conditioning in this experiment. A detailed description of the dataset and the corresponding adaptations is provided in the supplementary material, with additional qualitative results.

Despite these operator-level changes, IFIN outperforms purely data-driven baselines [15, 52] and a physics-initialized baseline that applies a single back-propagation step followed by NAFNet refinement (NAFNet+) (Tab. 3). This result indicates that the interleaving mechanism is transferable: when the forward and inverse operators are swapped to match a different imaging physics, the same architectural principle continues to provide effective guidance beyond deconvolution.

Table 4: Component and PSF-field ablation on DiffuserCam, plus learned-vs.-calibrated PSFs on MWNet ($k=9$).

Setting	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$
<i>Component controls — DiffuserCam ($k=1$)</i>			
Identity FSO/ISO	24.674	0.255	0.800
FSO-only (w/o ISO)	27.123	0.223	0.833
ISO-only (w/o FSO)	28.711	0.185	0.882
ConvG instead of RB	29.014	0.184	0.882
w/o 2D regularizer	28.880	0.186	0.881
w/o initial ISO	29.443	0.176	0.891
IFIN ($k=1$)	29.535	0.179	0.892
<i>PSF-field scaling — DiffuserCam</i>			
IFIN ($k=4$)	29.612	0.177	0.893
IFIN ($k=9$)	29.751	0.176	0.893
IFIN ($k=16$)	29.862	0.174	0.893
<i>Learned vs. calibrated PSF — MWNet ($k=9$)</i>			
Frozen calibrated PSFs	30.350	0.183	0.857
Learned PSF field	31.083	0.175	0.866

Table 5: Computational cost on DiffuserCam (batch size 1, RTX A6000): parameters, FLOPs, peak VRAM, and inference time.

Method	Params (M)	FLOPs (G)	VRAM (GB)	Time (ms)
NAFNet	29.2	32.4	0.257	24.65
MWDNs	22.1	106.4	0.422	12.81
UPDN	3.7	24.5	0.326	60.66
LensNet	31.6	147.2	0.529	43.00
MoDL	8.5	119.1	0.202	36.60
IFIN ($k=1$)	18.2	55.0	0.464	142.35
IFIN ($k=4$)	66.2	62.4	1.257	156.33
IFIN ($k=9$)	155.9	82.4	2.614	180.37
IFIN ($k=16$)	301.5	126.9	4.590	215.97

Ablation Study and Computational Cost. We isolate each component and quantify efficiency on DiffuserCam (Tab. 4 and Tab. 5; full ablations in the supplementary material). Replacing both operators with identities keeps the dual-stream paths but removes the physics, and the large drop shows that the gain originates in the forward-inverse coupling, not the extra pathways or parameters. Retaining a single direction (*FSO-only* or *ISO-only*) recovers only part of the gap, so the operators supply *complementary* cues: measurement-domain consistency from one and image-domain recoverability from the other.

The other controls locate where this benefit arises. Removing the learnable 2D regularizer or replacing the refinement block (RB) with a plain convolutional gate (ConvG) lowers quality, so the physics operators pay off only when paired with a learnable stabilizer and enough capacity; dropping the initial ISO barely changes the result; the initial ISO mainly helps stabilize training. Scaling the PSF field ($k=1 \rightarrow 16$) yields consistent but saturating gains, since a few well-placed kernels already capture most of the shift variance, and a learned field surpasses frozen calibrated PSFs, indicating that end-to-end refinement absorbs residual calibration mismatch.

These gains do not stem from scale alone: IFIN ($k=1$) already surpasses every baseline at comparable parameters, FLOPs, and memory, so the improvement is architectural. Its overhead is the latency of the padded FFT-based FSO/ISO invoked at every block, making $k=1$ a competitive low-cost operating point and larger k a deliberate accuracy-for-compute trade-off under strong shift variance.

6 Conclusion

We present **IFIN**, an integrated forward-inverse reconstruction architecture for lensless imaging that interleaves differentiable forward projections with learn-

able inverse updates across an encoder-decoder hierarchy. IFIN maintains coupled measurement- and image-domain streams and reintroduces measurement-consistency cues throughout decoding, letting the network refine image features while repeatedly checking their consistency under the imaging model. Furthermore, IFIN handles spatially varying systems and imperfect calibration using learnable PSF fields, allowing a single backbone to adapt to off-axis degradation that no stationary kernel can capture.

While IFIN achieves new state-of-the-art performance across diverse benchmarks (including our newly proposed WiderCam dataset) and shows promise in broader inverse problems, challenges remain. On the practical side, shift-variant systems with extended PSFs incur high per-block FFT overhead. The single-kernel ($k=1$) setting already offers a strong low-cost operating point, while denser PSF fields are reserved for stronger shift variance. IFIN also relies on a sufficiently stable forward model, as severe saturation or large geometry changes can erase information that no reconstructor can recover. We provide a comprehensive discussion of the generality, potential extensions, and limitations of our method in the supplementary material. Taken together, our results suggest that forward-inverse integration offers a practical template for designing reconstruction architectures in computational imaging systems.

Code and Data Availability

Code, supplementary material, and the WiderCam dataset are available at <https://iilab.io/IFIN/>.

Acknowledgements

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (RS-2026-25495603); by the Technology Innovation Program (IRIS number: RS-2024-00419426, Development of light-electron beam based measurement and analysis instrument technologies for advanced packaging) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea); and by the Creative-Pioneering Researchers Program through Seoul National University.

References

1. 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 (2017)
2. 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)
3. Asif, M.S., Ayremlou, A., Sankaranarayanan, A., Veeraraghavan, A., Baraniuk, R.G.: Flatcam: Thin, lensless cameras using coded aperture and computation. *IEEE Transactions on Computational Imaging* **3**(3), 384–397 (2016)

4. Bae, D., Jung, J., Baek, N., Lee, S.A.: Lensless imaging with an end-to-end deep neural network. In: 2020 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia). pp. 1–5. IEEE (2020)
5. Baek, N., Lee, Y., Kim, T., Jung, J., Lee, S.A.: Lensless polarization camera for single-shot full-stokes imaging. *APL Photonics* **7**(11) (2022)
6. Bagadthey, D., Prabhu, S., Khan, S.S., Fredrick, D.T., Boominathan, V., Veeraraghavan, A., Mitra, K.: Flatnet3d: intensity and absolute depth from single-shot lensless capture. *Journal of the Optical Society of America A* **39**(10), 1903–1912 (2022)
7. Bai, J., Yin, Y., Dong, Y., Zhang, X., Pun, C.M., Chen, X.: Lensnet: An end-to-end learning framework for empirical point spread function modeling and lensless imaging reconstruction. *arXiv* 2025. *arXiv preprint arXiv:2505.01755* (2025)
8. Bertero, M., Boccacci, P., De Mol, C.: Introduction to inverse problems in imaging. CRC press (2021)
9. Bezzam, E., Perron, Y., Vetterli, M.: Towards robust and generalizable lensless imaging with modular learned reconstruction. *IEEE Transactions on Computational Imaging* **11**, 213–227 (2025)
10. Boominathan, V., Adams, J.K., Robinson, J.T., Veeraraghavan, A.: Phlatcam: Designed phase-mask based thin lensless camera. *IEEE transactions on pattern analysis and machine intelligence* **42**(7), 1618–1629 (2020)
11. Booth, M.J.: Adaptive optical microscopy: the ongoing quest for a perfect image. *Light: Science & Applications* **3**(4), e165–e165 (2014)
12. 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)
13. 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)
14. Chen, D., Davies, M.E.: Deep decomposition learning for inverse imaging problems. In: *European Conference on Computer Vision*. pp. 510–526. Springer (2020)
15. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: *European conference on computer vision*. pp. 17–33. Springer (2022)
16. Cover, T.M.: Elements of information theory. John Wiley & Sons (1999)
17. Dong, J., Roth, S., Schiele, B.: Dwdn: Deep wiener deconvolution network for non-blind image deblurring. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(12), 9960–9976 (2021)
18. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020)
19. Faccio, D., Velten, A., Wetzstein, G.: Non-line-of-sight imaging. *Nature Reviews Physics* **2**(6), 318–327 (2020)
20. Ge, H., Feng, Q., Jia, H., Li, X., Yin, X., Zhou, Y., Yang, J., Li, K.: 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)
21. Henry, C., Asif, M.S., Li, Z.: 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)

22. Khan, S.S., Sundar, V., Boominathan, V., Veeraraghavan, A., Mitra, K.: Flatnet: Towards photorealistic scene reconstruction from lensless measurements. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(4), 1934–1948 (2020)
23. Kim, T., Lee, K.C., Lee, K., Baek, N., Jung, J., Kim, E., Park, B., Ha, J., Kim, K.Y., Seo, Y.S., et al.: High-speed lensless eye tracker for microsaccade measurement. In: *SPIE Advanced Biophotonics Conference (SPIE ABC 2023)*. vol. 13076, pp. 38–44. SPIE (2024)
24. Kingshott, O., Antipa, N., Bostan, E., Akşit, K.: Unrolled primal-dual networks for lensless cameras. *Optics Express* **30**(26), 46324–46335 (2022)
25. Lee, K.C., Bae, J., Baek, N., Jung, J., Park, W., Lee, S.A.: Design and single-shot fabrication of lensless cameras with arbitrary point spread functions. *Optica* **10**(1), 72–80 (2023)
26. Li, Y., Li, Z., Chen, K., Guo, Y., Rao, C.: Mwdns: reconstruction in multi-scale feature spaces for lensless imaging. *Optics Express* **31**(23), 39088–39101 (2023)
27. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: *ICCV* (2021)
28. Lucy, L.B.: An iterative technique for the rectification of observed distributions. *Astronomical Journal*, Vol. 79, p. 745 (1974) **79**, 745 (1974)
29. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. In: *NeurIPS* (2016)
30. Miao, J., Ishikawa, T., Robinson, I.K., Murnane, M.M.: Beyond crystallography: Diffractive imaging using coherent x-ray light sources. *Science* **348**(6234), 530–535 (2015)
31. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a “completely blind” image quality analyzer. *IEEE Signal processing letters* **20**(3), 209–212 (2012)
32. Monakhova, K., Tran, V., Kuo, G., Waller, L.: Untrained networks for compressive lensless photography. *Optics Express* **29**(13), 20913–20929 (2021)
33. Monakhova, K., Yanny, K., Aggarwal, N., Waller, L.: Spectral diffusercam: lensless snapshot hyperspectral imaging with a spectral filter array. *Optica* **7**(10), 1298–1307 (2020)
34. 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)
35. Pan, X., Chen, X., Takeyama, S., Yamaguchi, M.: Image reconstruction with transformer for mask-based lensless imaging. *Optics Letters* **47**(7), 1843–1846 (2022)
36. Pavani, S.R.P., Thompson, M.A., Biteen, J.S., Lord, S.J., Liu, N., Twieg, R.J., Piestun, R., Moerner, W.E.: 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)
37. 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)
38. Richardson, W.H.: Bayesian-based iterative method of image restoration. *Journal of the optical society of America* **62**(1), 55–59 (1972)
39. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention*. pp. 234–241. Springer (2015)
40. Sahoo, S.K., Tang, D., Dang, C.: Single-shot multispectral imaging with a monochromatic camera. *Optica* **4**(10), 1209–1213 (2017)

41. Satat, G., Tancik, M., Gupta, O., Heshmat, B., Raskar, R.: Object classification through scattering media with deep learning on time resolved measurement. *Optics express* **25**(15), 17466–17479 (2017)
42. Thiébaud, É., Denis, L., Soulez, F., Mourya, R.: Spatially variant psf modeling and image deblurring. In: *Adaptive Optics Systems V*. vol. 9909, pp. 2211–2220. SPIE (2016)
43. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Deep image prior. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 9446–9454 (2018)
44. Wang, F., Bian, Y., Wang, H., Lyu, M., Pedrini, G., Osten, W., Barbastathis, G., Situ, G.: Phase imaging with an untrained neural network. *Light: Science & Applications* **9**(1), 77 (2020)
45. Wiener, N.: *Extrapolation, interpolation, and smoothing of stationary time series*. The MIT press (1964)
46. Xiangjun, Y., Yue, H.: Reveal object in lensless photography via region gaze and amplification. In: *The Thirteenth International Conference on Learning Representations* (2025)
47. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: Maniqa: Multi-dimension attention network for no-reference image quality assessment. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 1191–1200 (2022)
48. Yanny, K., Antipa, N., Liberti, W., Dehaeck, S., Monakhova, K., Liu, F.L., Shen, K., Ng, R., Waller, L.: Miniscope3d: optimized single-shot miniature 3d fluorescence microscopy. *Light: Science & Applications* **9**(1), 171 (2020)
49. Yanny, K., Monakhova, K., Shuai, R.W., Waller, L.: Deep learning for fast spatially varying deconvolution. *Optica* **9**(1), 96–99 (2022)
50. Yoon, S., Kim, M., Jang, M., Choi, Y., Choi, W., Kang, S., Choi, W.: Deep optical imaging within complex scattering media. *Nature Reviews Physics* **2**(3), 141–158 (2020)
51. 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)
52. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 286–301 (2018)