LensNet: An End-to-End Learning Framework for Empirical Point Spread Function Modeling and Lensless Imaging Reconstruction

Jiesong Bai[1,2], **Yuhao Yin**[2], **Yihang Dong**[1], **Xiaofeng Zhang**[3], **Chi-Man Pun**[1], **Xuhang Chen**[1,4]
[1]Faculty of Science and Technology, University of Macau
[2]Shanghai University
[3]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University
[4]School of Computer Science and Engineering, Huizhou University
xuhangc@hzu.edu.cn

Abstract

Lensless imaging stands out as a promising alternative to conventional lens-based systems, particularly in scenarios demanding ultracompact form factors and cost-effective architectures. However, such systems are fundamentally governed by the Point Spread Function (PSF), which dictates how a point source contributes to the final captured signal. Traditional lensless techniques often require explicit calibrations and extensive pre-processing, relying on static or approximate PSF models. These rigid strategies can result in limited adaptability to real-world challenges, including noise, system imperfections, and dynamic scene variations, thus impeding high-fidelity reconstruction. In this paper, we propose LensNet, an end-to-end deep learning framework that integrates spatial-domain and frequency-domain representations in a unified pipeline. Central to our approach is a learnable Coded Mask Simulator (CMS) that enables dynamic, data-driven estimation of the PSF during training, effectively mitigating the shortcomings of fixed or sparsely calibrated kernels. By embedding a Wiener filtering component, LensNet refines global structure and restores fine-scale details, thus alleviating the dependency on multiple handcrafted pre-processing steps. Extensive experiments demonstrate LensNet's robust performance and superior reconstruction quality compared to state-of-the-art methods, particularly in preserving high-frequency details and attenuating noise. The proposed framework establishes a novel convergence between physics-based modeling and data-driven learning, paving the way for more accurate, flexible, and practical lensless imaging solutions for applications ranging from miniature sensors to medical diagnostics. The link of code is https://github.com/baijiesong/Lensnet.

1 Introduction

Conventional imaging systems that rely on lenses are inherently constrained by factors such as cost, weight, and physical dimensions, which limit their potential for miniaturization [Li *et al.*, 2024a; Guo *et al.*, 2025a; Li *et al.*, 2023b; Guo *et al.*, 2024; Guo *et al.*, 2025b; Zhou *et al.*, 2024; Chen *et al.*, 2023a]. In contrast, lensless imaging avoids the use of lenses altogether and instead uses computational algorithms to reconstruct scenes. This lens-free approach not only lowers production costs but also reduces device thickness, supporting the development of ultra-thin imagers. Existing methods frequently employ coded apertures [Asif *et al.*, 2017] or phase masks [Antipa *et al.*, 2017; Boominathan *et al.*, 2020] near image sensors to modulate incoming light. These compact and low-cost solutions are well suited for various applications, including miniature sensors and medical imaging. Without lenses, acquired signals become highly multiplexed and require sophisticated computational techniques [Wiener, 1949] for accurate reconstruction. Central to these techniques is the Point Spread Function (PSF), which describes how the system responds to a point source and significantly influences imaging performance.

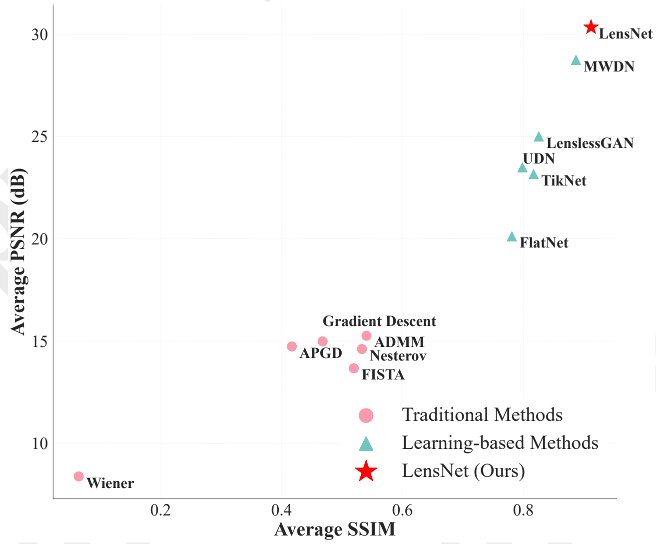

Figure 1: Performance comparison on the MWDNs dataset and the DiffuserCam dataset in terms of PSNR and SSIM.

Despite its pivotal role, traditional model-based methods [Monakhova *et al.*, 2019; Asif *et al.*, 2015; Tanida *et al.*,

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

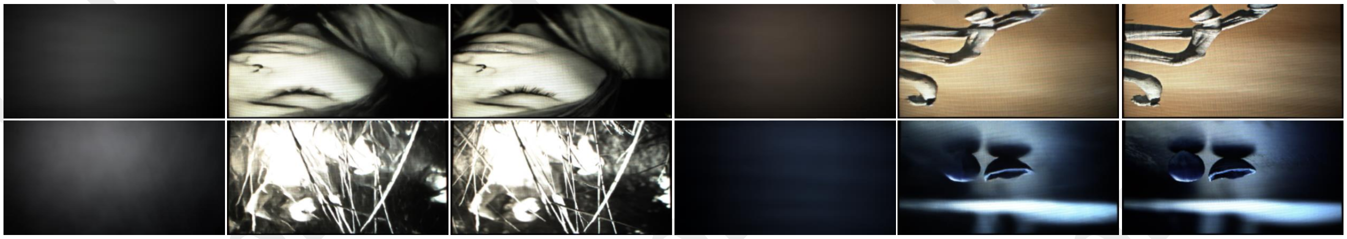

Figure 2: Visualized results on the DiffuserCam dataset. Each image shows the input image, the LensNet reconstruction result, and the corresponding ground truth.

2001] often rely on static or calibrated PSF models, which struggle to accommodate real-world challenges such as system imperfections and dynamic scene variations. In classic lensless imaging, the PSF is typically pre-calibrated or inferred through fixed models, leading to rigid assumptions that cannot adapt to practical scenarios. Factors such as changes in lighting, sensor noise, and minor misalignments cause deviations between the assumed PSF and the true PSF, ultimately degrading render quality. Furthermore, the large support of the PSF in lensless imaging presents two major challenges. First, the expanded linear equations resulting from a broad PSF footprint are computationally demanding to store and invert, leading to significant memory and processing overhead. Second, this global multiplexing complicates reconstructions as local image details are intertwined with the global signal. This issue becomes especially detrimental under low-light or noisy settings, where the recovery of fine-scale features is critical, yet hard to achieve.

Recent advances in deep learning [Liu *et al.*, 2023a; Zeng and Lam, 2021; Li *et al.*, 2024b; Zhu *et al.*, 2024b; Liu *et al.*, 2023b; Yao *et al.*, 2024; Pan *et al.*, 2024; Chen *et al.*, 2025; Chen *et al.*, 2023b; Luo *et al.*, 2024; Li *et al.*, 2023c; Chen *et al.*, 2024; Yang *et al.*, 2025; Huo *et al.*, 2024; Zhu *et al.*, 2024a] combine traditional imaging physics with data-driven neural networks and have achieved substantial progress in reconstructing scenes from heavily multiplexed signals. For example, FlatNet [Khan *et al.*, 2020] uses paired training data to improve image quality but encounters difficulties in recovering high-frequency details. Similarly, PhoCoLens [Cai *et al.*, 2024] employs a two-stage pipeline that shows promising performance, yet may blur high-frequency components during its diffusion procedure.

To address these shortcomings, we propose *LensNet*, an end-to-end deep learning framework that effectively fuses spatial and frequency domain information for lensless image reconstruction. Our framework adopts an encoder–decoder architecture, where the encoder features a novel *Coded Mask Simulator (CMS)* module that we design to capture the optical path and raster encoding properties of the lensless imaging process. By modeling the global and multi-scale characteristics of a learnable PSF, this module adapts to diverse imaging conditions without requiring explicit PSF calibration. Furthermore, we incorporate Wiener filtering [Wiener, 1949] into our network to refine reconstructions via frequency-domain deconvolution and PSF correction, thereby enhancing noise suppression and edge fidelity. By integrating these customized modules, LensNet takes advantage of both spatial and frequency domain cues to produce robust, high-resolution reconstructions while maintaining global consistency, as showed in Fig. 2. Extensive experiments confirm the efficacy and reliability of our design. Our main contributions are as follows.

- We propose a novel end-to-end deep learning framework for lensless imaging that dynamically captures multi-scale features in both spatial and frequency domains, substantially improving the fidelity and accuracy of image reconstructions over conventional methods.
- We introduce a state-of-the-art Coded Mask Simulator that precisely learns the true behavior of the point spread function (PSF), facilitating superior recovery performance across diverse lensless imaging scenarios.
- Through comprehensive quantitative and visual evaluations, we demonstrate that our proposed LensNet outperforms existing approaches on standard benchmark datasets.

2 Related Work

2.1 Traditional Lensless Imaging

Traditional lensless imaging methods commonly rely on optical encoding techniques. Coded aperture imaging [Asif *et al.*, 2017] employs a structured mask or aperture to encode incident light, while phase retrieval [Antipa *et al.*, 2017; Boominathan *et al.*, 2020] reconstructs both the amplitude and phase of the light field from intensity measurements. Although these approaches have shown success in specific scenarios, they often require complex optical setups and computationally expensive reconstruction algorithms that are prone to noise and challenging imaging conditions.

In addition to encoding-based methods, iterative optimization methodologies such as FISTA [Beck and Teboulle, 2009a] and ADMM [Wang *et al.*, 2019] are widely adopted to address inverse problems in lensless imaging. FISTA accelerates sparse signal recovery, while ADMM efficiently solves convex optimization problems by decomposing them into simpler sub-problems. Moreover, ADMM with total variation regularization [Wahlberg *et al.*, 2012] helps to preserve edges and reduce noise. Despite the effectiveness of these methods, they require careful parameter tuning and remain computationally expensive, particularly under noisy measurements or complex optical systems.

2.2 Learning-Based Lensless Imaging

Recent deep learning approaches [Zheng *et al.*, 2024; Liu *et al.*, 2024; Monakhova *et al.*, 2019; Kingshott *et al.*, 2022; Liu *et al.*, 2023c] have capitalized on the representational power of neural networks to model complex forward and reverse imaging processes directly from data. Multiple studies [Cai *et al.*, 2024; Khan *et al.*, 2019; Li *et al.*, 2023a] have validated the effectiveness of these learning-based methods. For example, TikNet and FlatNet [Khan *et al.*, 2020] use paired training data to improve image quality but struggle to preserve high-frequency details. Similarly, PhoCoLens [Cai *et al.*, 2024] employs a two-stage processing framework that achieves favorable results; however, its diffusion process blurs fine structures, requiring iterative refinements. [Rego *et al.*, 2021] proposes a lensless imaging reconstruction method based on GAN, which achieves robust reconstruction of images captured by multiple cameras without requiring explicit PSF knowledge by jointly performing image reconstruction and PSF estimation.

Despite these advances, several challenges remain. Precise recovery of fine-scale image details, accommodating fixed or overly simplistic point spread functions (PSFs), and mitigating noise-induced degradation are significant hurdles. Traditional systems often rely on static PSF models that cannot adapt to variations in the scene or account for sensor imperfections, thus hindering image quality. Preserving high-frequency vision cues, such as sharp edges and detailed textures, remains also difficult. Our proposed imaging framework addresses these limitations by introducing a more flexible PSF modeling mechanism and improving the restoration of high-frequency details.

3 Methodology

In this section, we present a novel lensless imaging framework that jointly exploits both spatial and frequency domain information for high-fidelity image reconstruction. To address the limitations of existing lensless imaging systems, our proposed framework integrates multiple advanced modules specifically designed to overcome inefficiencies and inaccuracies commonly encountered in PSF-based approaches. The remainder of this section describes the overarching network architecture and its constituent modules in detail.

3.1 Preliminary

In optics, the Point Spread Function (PSF), denoted as $\mathrm{PSF}(x, y)$, characterizes how an imaging system distributes a point source of light. Technically, $\mathrm{PSF}(x, y)$ is a two-dimensional or three-dimensional function representing the output signal when a point source is introduced at the input. Coded masks, widely employed as optical encoders, are constructed with specific patterns (e.g., horizontal lines, vertical lines, grids, or random structures) to modulate incident light. By imposing spatial modulation on the optical signal, these masks effectively encode essential information about the scene. In lensless imaging systems, the particular arrangement of the coded mask plays a pivotal role in the definition of $\mathrm{PSF}(x, y)$, thereby influencing the quality and fidelity of the reconstructed images. Hence, the design of the coded mask significantly affects the imaging performance of the system, determining how light is distributed and shaping critical attributes such as resolution, contrast, and image clarity.

The forward lensless imaging model can be mathematically represented as a linear transformation [Goodman, 2005]. For an input image, the sensor in a lensless system acquires a blurred observation:

$$I_{\text{measurement}}(x, y) = I_{\text{object}}(x, y) * \mathrm{PSF}(x, y) + \text{noise}(x, y), \tag{1}$$

where $I_{\text{object}}(x, y)$ is the original image, $\mathrm{PSF}(x, y)$ denotes the system's point spread function, $\text{noise}(x, y)$ is the sensor noise, and $I_{\text{measurement}}(x, y)$ is the recorded measurement.

By introducing spatial encoding through coded masks, lensless systems modulate the incoming light to create distinctive intensity distributions on the detector. Physically, the coded mask pattern directly affects $\mathrm{PSF}(x, y)$, which emerges in the measurement image via the convolution operation:

$$I_{\text{measurement}}(x, y) = I_{\text{object}}(x, y) * \mathrm{PSF}(x, y). \tag{2}$$

Hence, the mask pattern governs the distribution of spatial frequencies, the direction of light propagation, and the overall optical properties, all of which are intrinsically manifested in $I_{\text{measurement}}(x, y)$. Consequently, the choice of coded mask has a profound influence on both the image formation process and the achievable reconstruction quality.

3.2 Overall Architecture

The overall architecture of LensNet is illustrated in Fig. 3. The framework adopts an encoder-decoder structure, where the encoder processes multi-resolution spatial and frequency domain features in the first stage, and the decoder reconstructs the image in the second stage. We begin by outlining the LensNet framework and then provide a detailed description of its individual components.

Given the measurement image $I_{\text{measurement}}(x, y)$ measured by the sensor in a lensless imaging system, we simplify the analysis by momentarily assuming a noise-free condition. The goal of LensNet is to reconstruct $I_{\text{object}}(x, y)$ from $I_{\text{measurement}}(x, y)$ while satisfying the following equation:

$$I_{\text{measurement}}(x, y) = I_{\text{object}}(x, y) * \mathrm{PSF}(x, y). \tag{3}$$

In the encoder stage, we incorporate a Spatial Compression Module (SCM), which leverages Reconstruction Blocks (ReCB) as its primary feature transformation mechanism and employs gating strategies to enhance feature fusion. First, $I_{\text{measurement}}(x, y)$ is processed on multiple scales, extracting spatial context at varying resolutions. These features establish the foundation for reconstructing the spatial representation of the object.

To account for the coded mask distribution associated with the system PSF, we introduce a learnable Coded Mask Simulator (CMS). This module applies a dynamic channel attention mechanism to modulate light intensity across different regions, effectively modeling the coded mask's influence on the system's PSF. By learning the relationship between spatial inputs and the PSF, the CMS dynamically adapts to changing

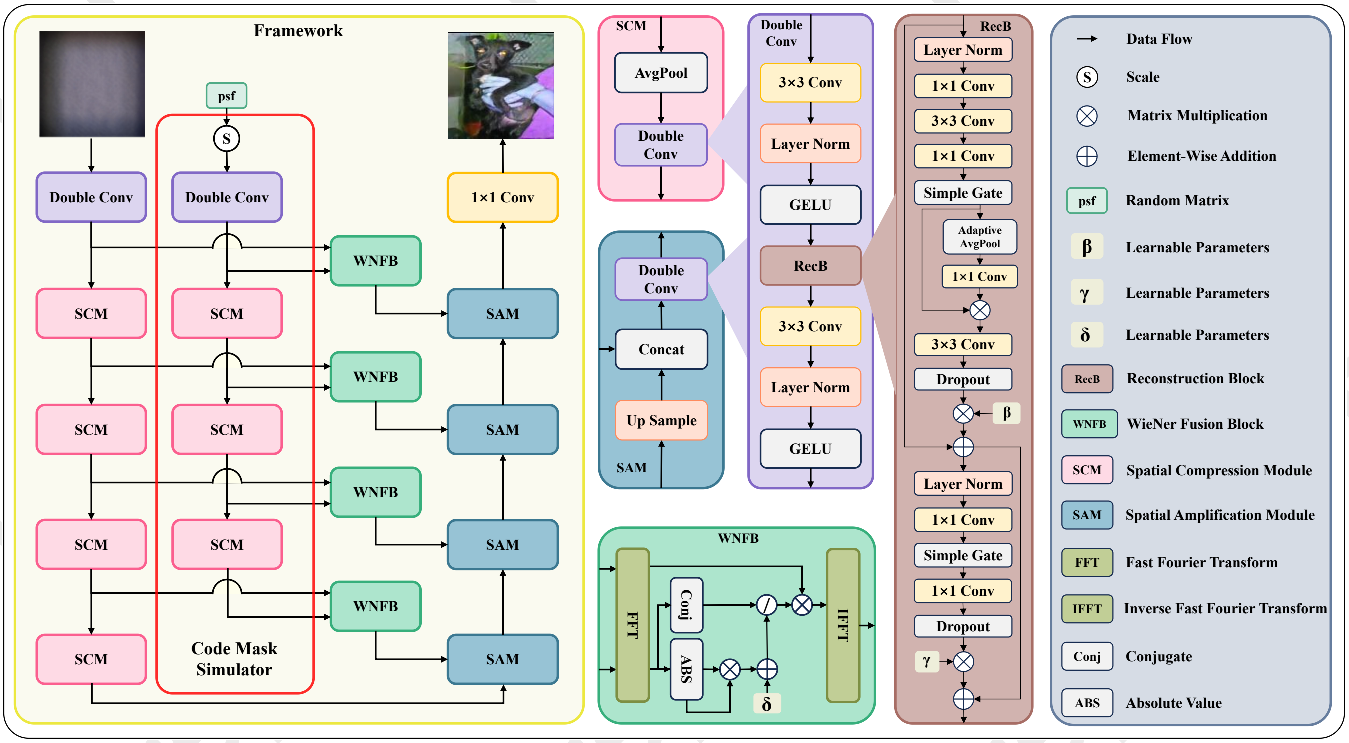

Figure 3: The figure illustrates the detailed structure of our model. The input image is measurement on the system and the result is clear scene image through our framework.

imaging conditions, improving the simulation of light intensity patterns. Its output, a distribution map that encodes high-level semantic cues, facilitates more accurate reconstruction of structures, especially in complex or dynamic scenarios where traditional fixed PSF methods often fail.

We further extract frequency domain features at each scale via the Wiener Fusion Block (WNFB). This block performs a Fast Fourier Transform (FFT) on both the spatial features and the output of the Coded Mask Simulator, and then computes the Wiener filter transfer function:

$$H(u,v) = \frac{\overline{\mathrm{PSF}(u,v)}}{|\,\mathrm{PSF}(u,v)|^2 + \delta}, \tag{4}$$

where $H(u,v)$ denotes the Wiener filter in the frequency domain, $\overline{\mathrm{PSF}(u,v)}$ is the complex conjugate of the PSF, and $|\,\mathrm{PSF}(u,v)|^2$ is the energy spectrum of the PSF.

Using this transfer function, the encoded frequency-domain signal $B(u,v)$ is preliminarily restored:

$$I_{\text{restored}}(u,v) = H(u,v) \cdot B(u,v). \tag{5}$$

In the decoder stage, we propose the Spatial Amplification Module (SAM) to integrate spatial and frequency domain information more comprehensively. The SAM upsamples low-resolution feature maps from deeper layers via bilinear interpolation and aligns them (through zero-padding) with higher-resolution encoder features. These feature maps are then concatenated to preserve both fine-grained detail and broader contextual cues. Serving as a multi-scale fusion mechanism, the SAM leverages information from both domains to produce high-quality reconstructions, ensuring that critical spatial details and global consistency are well-preserved.

3.3 Coded Mask Simulator

As discussed in Sec. 3.1, the physical level $\mathrm{PSF}(x,y)$ corresponds to the coded mask placed within the sensor. Traditional methods commonly adopt static or fixed kernels to approximate $\mathrm{PSF}(x,y)$. In contrast, we seek a more flexible approach that adaptively learns these coded patterns, resulting in a learnable PSF and minimizing constraints seen in conventional models.

To address this need, we propose the Coded Mask Simulator (CMS). By employing a channel attention mechanism, the CMS captures the intensity distribution features in $I_{\text{measurement}}(x,y)$, which encode the coded mask pattern. Consequently, the system's $\mathrm{PSF}(x,y)$ can be inferred from these learned mask distributions. This attention mechanism models directional inter-pixel relationships, enabling the network to deconvolve scattering and diffraction artifacts induced by the mask. More concretely, it identifies directional modulations in the image and adaptively recalibrates features according to attention weights, achieving a more robust partial or full recovery of occluded signals.

Formally, let $X \in \mathbb{R}^{N\times C\times H\times W}$ be the input feature map. We first apply global average pooling to aggregate spatial in-

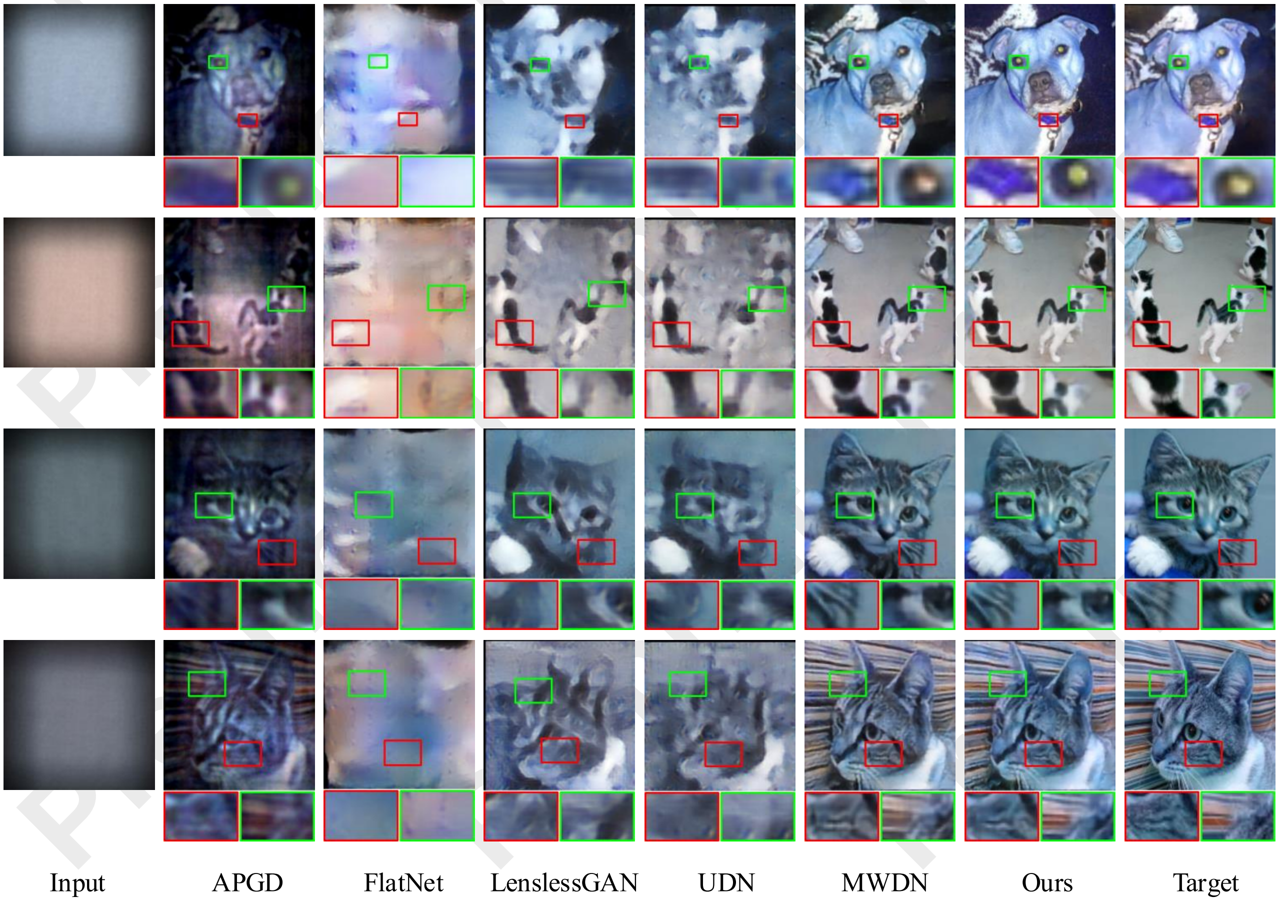

Figure 4: The above shows the qualitative experimental results on MWDNs Dataset. Our method effectively achieves high-quality reconstruction, comparing other methods.

formation, preserving only channel-level statistics:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{c,i,j}, \quad \forall c \in \{1, 2, \ldots, C\}. \tag{6}$$

Subsequently, a pointwise convolution adjusts the resulting channel weights. This operation, which is more efficient than deeper multilayer perceptrons, preserves channel-specific modulation dynamics:

$$s_c = \sigma(W \cdot z + b), \tag{7}$$

where $W \in \mathbb{R}^{C' \times C'}$ (with $C' = \frac{C}{2}$) is the 1×1 convolution weight matrix, σ is a (linear) activation function, and $s_c \in \mathbb{R}^{N \times C' \times 1 \times 1}$ are the channel attention weights.

In summary, the proposed Coded Mask Simulator models the physical coded mask by learning its intensity distribution and directional modulation patterns from the measurement image. This approach implicitly recovers $\mathrm{PSF}(x, y)$ and provides valuable semantic cues that guide subsequent reconstruction steps. The learned distribution fosters flexible, adaptive, and comprehensive recovery of scene information, overcoming the drawbacks associated with traditional fixed PSF kernels.

3.4 Objective Function

We train LensNet by jointly minimizing three complementary loss terms: the mean squared error (MSE), the structural similarity index measure (SSIM) and the perceptual image patch similarity (LPIPS). Formally, we define the overall loss function as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{MSE}} + 0.2 \cdot \mathcal{L}_{\mathrm{SSIM}} + 0.2 \cdot \mathcal{L}_{\mathrm{LPIPS}}, \tag{8}$$

where each loss provides a different perspective on reconstruction fidelity.

1. Mean Squared Error (MSE)

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{N} \sum_{i=1}^{N} \left\| \widehat{\mathbf{I}}_i - \mathbf{I}_i \right\|^2, \tag{9}$$

where $\widehat{\mathbf{I}}_i$ denotes the output image, and $\mathbf{I}_i$ represents the corresponding ground truth. By quantifying pixel-wise intensity differences, MSE imposes strong penalties for large deviations and steadily guides the network toward minimizing low-level reconstruction errors.

2. Structural Similarity Index Measure (SSIM)

$$\mathcal{L}_{\mathrm{SSIM}} = 1 - \mathrm{SSIM}(\widehat{\mathbf{I}}_i, \mathbf{I}_i), \tag{10}$$

Method	DiffuserCam			MWDNs		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Wiener [Wiener, 1949]	7.33	0.083	0.770	9.44	0.045	0.731
Vanilla Gradient Descent	13.27	0.432	0.585	16.70	0.503	0.429
Nesterov Gradient Descent [Nesterov, 1998]	12.16	0.394	0.518	17.06	0.671	0.362
FISTA [Beck and Teboulle, 2009b]	11.09	0.341	0.554	16.25	0.697	0.368
ADMM [Boyd *et al.*, 2011]	12.76	0.442	0.541	17.76	0.638	0.343
APGD [Li and Lin, 2015]	12.13	0.385	0.518	17.34	0.448	0.439
TikNet [Khan *et al.*, 2020]	19.75	0.720	0.221	26.57	0.913	0.075
FlatNet [Khan *et al.*, 2020]	21.16	0.720	0.231	19.08	0.841	0.178
LenslessGAN [Rego *et al.*, 2021]	22.51	0.737	0.193	27.49	0.913	0.077
UDN [Kingshott *et al.*, 2022]	20.00	0.688	0.250	26.98	0.908	0.081
MWDN [Li *et al.*, 2023a]	25.74	0.816	0.132	31.74	0.957	0.030
LensNet	**27.46**	**0.863**	**0.099**	**33.22**	**0.960**	**0.024**

Table 1: Performance comparison on two datasets: DiffuserCam and MWDNs. Metrics include PSNR (dB), SSIM, and LPIPS. Best results are highlighted in red and bold.

This term evaluates perceptual quality by examining local image structures. SSIM is particularly effective for preserving global consistency and fine structural details, ensuring that the reconstructed images align with human perception of image integrity.

3. Learned Perceptual Image Patch Similarity (LPIPS)

$$\mathcal{L}_{\mathrm{LPIPS}} = \mathrm{LPIPS}(\widehat{\mathbf{I}}_i, \mathbf{I}_i). \tag{11}$$

LPIPS leverages deep neural network features to evaluate perceptual similarity. Unlike MSE or SSIM, which target lower-level and local structure information, LPIPS operates on higher-level representations, promoting reconstructions that faithfully capture the semantic and perceptual nuances of the scene.

By combining these three loss components, our objective function balances low-level fidelity, structural coherence, and perceptual plausibility. Consequently, this strategy leads to more accurate and visually coherent reconstructions in lensless imaging.

4 Experiments

4.1 Dataset and Preprocessing

The DiffuserCam dataset [Antipa *et al.*, 2017; Monakhova *et al.*, 2019] comprises 25,000 pairs of images captured simultaneously by a standard lensed camera (serving as ground truth) and a diffuser-based lensless camera. The raw images, originally at 1080×1920 pixels, are resized to 480×270 pixels according to previous methods [Khan *et al.*, 2020; Kingshott *et al.*, 2022; Li *et al.*, 2023a; Rego *et al.*, 2021]. Of these 25,000 paired images, 24,000 are allocated for training and 1,000 for testing.

Similarly, the MWDNs dataset [Li *et al.*, 2023a] consists of 25,000 images displayed on a 27-inch Dell monitor, sourced from the Dogs vs. Cats dataset on Kaggle. Original 1280×1280 images undergo the official preprocessing steps (zero-padding and resizing) to obtain 320×320 pixels. This standard preprocessing pipeline ensures consistent input dimensions across both datasets, facilitating robust and fair comparisons of lensless imaging strategies.

4.2 Implementation Details

We implemented our method in PyTorch and conducted experiments on 2 NVIDIA RTX 3090 GPUs running Ubuntu. All images were used at their preprocessed resolutions without further downsampling. We set the batch size to 16 for both datasets and applied standard data augmentation strategies, such as random rotation and flipping. For optimization, we employed the Adam optimizer.

4.3 Evaluation Metrics

We evaluated the performance of our lensless imaging approach using three widely adopted image quality metrics: PSNR, SSIM, and LPIPS. PSNR quantifies the level of noise reduction, while SSIM measures the preservation of structural and textural fidelity. LPIPS also assesses perceptual similarity by comparing reconstructed images with the ground truth according to human visual perception. Higher PSNR and SSIM values indicate better image quality and structural integrity, while lower LPIPS values suggest improved perceptual fidelity and preservation of fine details.

4.4 Comparisons with State-of-the-art Methods

In this section, we compare our proposed LensNet with traditional and learning-based methods to assess reconstruction fidelity, robustness, and overall performance in the DiffuserCam and MWDNs datasets.

Quantitative Comparisons

As summarized in Table 1, traditional algorithms such as Wiener, FISTA, ADMM, and APGD produce relatively low numerical performance when applied to highly multiplexed signals observed in lensless imaging. For instance, in the DiffuserCam dataset, these methods yield PSNR values typically between 7.33 dB and 13.27 dB, indicating their limited ability to invert complex degradations. Similarly, iterative optimization algorithms (Vanilla Gradient Descent and Nesterov Gradient Descent) exhibit difficulties in converging to high-quality solutions, particularly in low-light or noisy conditions.

On the other hand, learning-based approaches demonstrate noticeable improvements in reconstruction quality. FlatNet [Khan *et al.*, 2020] achieves PSNR values of 21.16 dB (DiffuserCam) and 19.08 dB (MWDN). Meanwhile, UDN [Kingshott *et al.*, 2022] and TikNet [Khan *et al.*, 2020] tend to preserve structural signals better than traditional methods, with UDN achieving 26.98 dB in PSNR on MWDNs. MWDN [Li *et al.*, 2023a], which uses a multi-scale wavelet decomposition strategy, further improves both PSNR and SSIM.

However, none of the previous methods dominate in every metric. Although UDN excels at preserving edges, it can still miss critical high-frequency texture in challenging illumination patterns. Similarly, though MWDN shows promising fine detail recovery, its susceptibility to high noise levels sometimes results in artifacts or mild oversmoothing under extreme conditions. In contrast, our proposed LensNet consistently outperforms all competing methods across the board, establishing new state-of-the-art results (PSNR: 27.46 dB in DiffuserCam and 33.22 dB in MWDN) alongside substantially lower LPIPS values. And Fig. 1 shows more intuitively.

Qualitative Comparisons

Beyond numerical metrics, we conduct an in-depth qualitative assessment to illustrate how LensNet handles subtle textures, fine details, and complex lighting conditions. Figure 4 presents representative reconstructions from the MWDNs datasets, comparing our method with both traditional and CNN-based baselines, with extended comparisons provided in the Supplementary Material.

As shown in Fig. 4, APGD and other traditional solvers often exhibit visibly blurred regions with pronounced ringing artifacts. FlatNet [Khan *et al.*, 2020] significantly reduces these artifacts, but introduces oversmoothing, obscuring key high-frequency details such as hair strands or fine-textured fabric. LenslessGAN [Rego *et al.*, 2021] improves the realism of certain regions; however, it can occasionally generate hallucinated details or struggle with complex boundaries under minimal SNR settings.

By contrast, LensNet consistently recovers sharper edges and more faithful textures, demonstrating well-preserved contours around objects and a closer match to the true intensity profile. As highlighted in the red and green bounding boxes of Fig. 4, LensNet excels at reconstructing intricate structures like an animal's fur or subtle shading gradients in natural scenes.

Methods	LensNet		
	PSNR↑	SSIM↑	LPIPS↓
ThreeDown	25.68	0.910	0.084
w/o RecB	30.36	0.945	0.044
w original PSF	31.92	0.954	0.031
Ours	**33.22**	**0.960**	**0.024**

Table 2: Ablation Study on LensNet

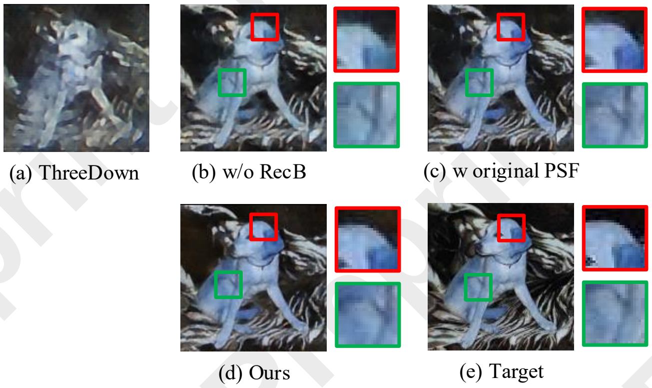

Figure 5: This figure presents the results of our ablation study, highlighting the impact of various components on model performance.

4.5 Ablation Study

As shown in Table 2, we conducted a detailed ablation study to verify the contributions of different components in our network. A reduced model with only three downsampling layers (ThreeDown) suffers from inadequate reconstruction quality, leading to notable losses in structural and textural details. When the ReconstructionBlock is removed (w/o RecB), the model still fails to preserve crucial local textures necessary for high-fidelity image reconstruction. Even when using the original point spread function (PSF) method, the performance did not exceed that of our learnable PSF design.

Illustrated in Fig. 5, the visual results underscore the importance of each architectural component. The ThreeDown variant fails to produce structurally coherent images, while omitting RecB degrades local texture preservation. In contrast, our full model—incorporating both the ReconstructionBlock and the learnable PSF—delivers results that are more faithful to the target images, retaining finer details and exhibiting overall higher visual quality. These observations confirm that each proposed module is crucial for achieving high-accuracy lensless imaging reconstructions.

5 Conclusion

In this study, we present LensNet, a novel deep learning framework designed for lensless imaging, which effectively integrates features in both the spatial and frequency domain to address the inherent challenges of lensless image reconstruction. By incorporating a learnable Coded Mask Simulator, LensNet dynamically models optical pathways and the Point Spread Function, enabling the precise and efficient reconstruction of complex scene details, including high-frequency components and fine textures. Our extensive experiments demonstrate that LensNet outperforms state-of-the-art methods, achieving superior image quality across multiple benchmark datasets. Moreover, its robustness in handling noise and variations in the PSF positions LensNet as a promising solution for practical applications in fields such as miniature sensors and medical diagnostics. This work not only enhances the capabilities of lensless imaging systems but also establishes a new standard for image reconstruction in lens-free environments.

Ethical Statement

There are no ethical issues.

Acknowledgments

This work was supported by the Science and Technology Development Fund, Macau SAR, under Grant 0141/2023/RIA2 and 0193/2023/RIA3.

Contribution Statement

This work was a collaborative effort by all contributing authors. Xuhang Chen, serving as the corresponding author, is responsible for all communications related to this manuscript.

References

[Antipa *et al.*, 2017] 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.

[Asif *et al.*, 2015] M Salman Asif, Ali Ayremlou, Ashok Veeraraghavan, Richard Baraniuk, and Aswin Sankaranarayanan. Flatcam: Replacing lenses with masks and computation. In *ICCV Workshop*, pages 663–666, 2015.

[Asif *et al.*, 2017] 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, 2017.

[Beck and Teboulle, 2009a] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[Beck and Teboulle, 2009b] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[Boominathan *et al.*, 2020] Vivek Boominathan, Jesse K Adams, Jacob T Robinson, and Ashok Veeraraghavan. Phlatcam: Designed phase-mask based thin lensless camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(7):1618–1629, 2020.

[Boyd *et al.*, 2011] 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.

[Cai *et al.*, 2024] Xin Cai, Zhiyuan You, Hailong Zhang, Wentao Liu, Jinwei Gu, and Tianfan Xue. Phocolens: Photorealistic and consistent reconstruction in lensless imaging. *arXiv*, 2024.

[Chen *et al.*, 2023a] Xuhang Chen, Xiaodong Cun, Chi-Man Pun, and Shuqiang Wang. Shadocnet: Learning spatial-aware tokens in transformer for document shadow removal. In *ICASSP*, pages 1–5, 2023.

[Chen *et al.*, 2023b] Xuhang Chen, Baiying Lei, Chi-Man Pun, and Shuqiang Wang. Brain diffuser: An end-to-end brain image to brain network pipeline. In *PRCV*, pages 16–26, 2023.

[Chen *et al.*, 2024] Xuhang Chen, Chi-Man Pun, and Shuqiang Wang. Medprompt: Cross-modal prompting for multi-task medical image translation. In *PRCV*, pages 61–75, 2024.

[Chen *et al.*, 2025] Guanli Chen, Guoheng Huang, Xiaochen Yuan, Xuhang Chen, Guo Zhong, and Chi-Man Pun. Cross-view geo-localization via learning correspondence semantic similarity knowledge. In *MMM*, pages 220–233, 2025.

[Goodman, 2005] Joseph W Goodman. *Introduction to Fourier optics*. Roberts and Company publishers, 2005.

[Guo *et al.*, 2024] Xiaojiao Guo, Xuhang Chen, Shenghong Luo, Shuqiang Wang, and Chi-Man Pun. Dual-hybrid attention network for specular highlight removal. In *ACM MM*, pages 10173–10181, 2024.

[Guo *et al.*, 2025a] Xiaojiao Guo, Xuhang Chen, Shuqiang Wang, and Chi-Man Pun. Underwater image restoration through a prior guided hybrid sense approach and extensive benchmark analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(5):4784–4800, 2025.

[Guo *et al.*, 2025b] Xiaojiao Guo, Yihang Dong, Xuhang Chen, Weiwen Chen, Zimeng Li, FuChen Zheng, and Chi-Man Pun. Underwater image restoration via polymorphic large kernel cnns. In *ICASSP*, pages 1–5, 2025.

[Huo *et al.*, 2024] Yejing Huo, Guoheng Huang, Lianglun Cheng, Jianbin He, Xuhang Chen, Xiaochen Yuan, Guo Zhong, and Chi-Man Pun. Iman: An adaptive network for robust npc mortality prediction with missing modalities. In *BIBM*, pages 2074–2079, 2024.

[Khan *et al.*, 2019] Salman S Khan, VR Adarsh, Vivek Boominathan, Jasper Tan, Ashok Veeraraghavan, and Kaushik Mitra. Towards photorealistic reconstruction of highly multiplexed lensless images. In *ICCV*, pages 7860–7869, 2019.

[Khan *et al.*, 2020] 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.

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

[Li and Lin, 2015] Huan Li and Zhouchen Lin. Accelerated proximal gradient methods for nonconvex programming. *NeurIPS*, 28, 2015.

[Li *et al.*, 2023a] 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.

[Li *et al.*, 2023b] Zinuo Li, Xuhang Chen, Chi-Man Pun, and Xiaodong Cun. High-resolution document shadow removal via a large-scale real-world dataset and a frequency-aware shadow erasing net. In *ICCV*, pages 12449–12458, 2023.

[Li *et al.*, 2023c] Zinuo Li, Xuhang Chen, Shuqiang Wang, and Chi-Man Pun. A large-scale film style dataset for learning multi-frequency driven film enhancement. In *IJCAI*, pages 1160–1168, 2023.

[Li *et al.*, 2024a] Mingxian Li, Hao Sun, Yingtie Lei, Xiaofeng Zhang, Yihang Dong, Yilin Zhou, Zimeng Li, and Xuhang Chen. High-fidelity document stain removal via a large-scale real-world dataset and a memory-augmented transformer. In *WACV*, pages 7614–7624, 2024.

[Li *et al.*, 2024b] Xiaohong Li, Guoheng Huang, Lianglun Cheng, Guo Zhong, Weihuang Liu, Xuhang Chen, and Muyan Cai. Cross-domain visual prompting with spatial proximity knowledge distillation for histological image classification. *Journal of Biomedical Informatics*, 158:104728, 2024.

[Liu *et al.*, 2023a] Muyuan Liu, Xiuqin Su, Xiaopeng Yao, Wei Hao, and Wenhua Zhu. Lensless image restoration based on multi-stage deep neural networks and pix2pix architecture. In *Photonics*, volume 10, page 1274, 2023.

[Liu *et al.*, 2023b] Weihuang Liu, Xiaodong Cun, Chi-Man Pun, Menghan Xia, Yong Zhang, and Jue Wang. Coordfill: Efficient high-resolution image inpainting via parameterized coordinate querying. In *AAAI*, volume 37, pages 1746–1754, 2023.

[Liu *et al.*, 2023c] Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *CVPR*, pages 19434–19445, 2023.

[Liu *et al.*, 2024] Weihuang Liu, Xi Shen, Haolun Li, Xiuli Bi, Bo Liu, Chi-Man Pun, and Xiaodong Cun. Depth-aware test-time training for zero-shot video object segmentation. In *CVPR*, pages 19218–19227, 2024.

[Luo *et al.*, 2024] Shenghong Luo, Xuhang Chen, Weiwen Chen, Zinuo Li, Shuqiang Wang, and Chi-Man Pun. Devignet: High-resolution vignetting removal via a dual aggregated fusion transformer with adaptive channel expansion. In *AAAI*, pages 4000–4008, 2024.

[Monakhova *et al.*, 2019] 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.

[Nesterov, 1998] Yurii Nesterov. Introductory lectures on convex programming volume i: Basic course. *Lecture notes*, 3(4):5, 1998.

[Pan *et al.*, 2024] Xiaochao Pan, Jiawei Yao, Hongrui Kou, Tong Wu, and Canran Xiao. Harmonicnerf: Geometry-informed synthetic view augmentation for 3d scene reconstruction in driving scenarios. In *ACM MM*, pages 5987–5996, 2024.

[Rego *et al.*, 2021] Joshua D Rego, Karthik Kulkarni, and Suren Jayasuriya. Robust lensless image reconstruction via psf estimation. In *WACV*, pages 403–412, 2021.

[Tanida *et al.*, 2001] Jun Tanida, Tomoya Kumagai, Kenji Yamada, Shigehiro Miyatake, Kouichi Ishida, Takashi Morimoto, Noriyuki Kondou, Daisuke Miyazaki, and Yoshiki Ichioka. Thin observation module by bound optics (tombo): concept and experimental verification. *Applied optics*, 40(11):1806–1813, 2001.

[Wahlberg *et al.*, 2012] Bo Wahlberg, Stephen Boyd, Mariette Annergren, and Yang Wang. An admm algorithm for a class of total variation regularized estimation problems. *IFAC Proceedings Volumes*, 45(16):83–88, 2012.

[Wang *et al.*, 2019] Yu Wang, Wotao Yin, and Jinshan Zeng. Global convergence of admm in nonconvex nonsmooth optimization. *Journal of Scientific Computing*, 78:29–63, 2019.

[Wiener, 1949] Norbert Wiener. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. The MIT press, 1949.

[Yang *et al.*, 2025] Jietao Yang, Guoheng Huang, Yanzhang Luo, Xiaofeng Zhang, Xiaochen Yuan, Xuhang Chen, Chi-Man Pun, and Mu-Yan Cai. Tcia: A transformer-cnn model with illumination adaptation for enhancing cell image saliency and contrast. *IEEE Transactions on Instrumentation and Measurement*, 74:1–15, 2025.

[Yao *et al.*, 2024] Jiawei Yao, Chuming Li, and Canran Xiao. Swift sampler: Efficient learning of sampler by 10 parameters. *arXiv preprint arXiv:2410.05578*, 2024.

[Zeng and Lam, 2021] Tianjiao Zeng and Edmund Y. Lam. Robust reconstruction with deep learning to handle model mismatch in lensless imaging. *IEEE Transactions on Computational Imaging*, 7:1080–1092, 2021.

[Zheng *et al.*, 2024] Fuchen Zheng, Xuhang Chen, Weihuang Liu, Haolun Li, Yingtie Lei, Jiahui He, Chi-Man Pun, and Shounjun Zhou. Smaformer: Synergistic multi-attention transformer for medical image segmentation. In *BIBM*, 2024.

[Zhou *et al.*, 2024] Ziyang Zhou, Yingtie Lei, Xuhang Chen, Shenghong Luo, Wenjun Zhang, Chi-Man Pun, and Zhen Wang. Docdeshadower: Frequency-aware transformer for document shadow removal. In *SMC*, pages 2468–2473, 2024.

[Zhu *et al.*, 2024a] Dingzhou Zhu, Guoheng Huang, Xiaochen Yuan, Xuhang Chen, Guo Zhong, Chi-Man Pun, and Jie Deng. Faqnet: Frequency-aware quaternion network for endoscopic highlight removal. In *BIBM*, pages 1408–1413, 2024.

[Zhu *et al.*, 2024b] Leyi Zhu, Weihuang Liu, Xinyi Chen, Zimeng Li, Xuhang Chen, Zhen Wang, and Chi-Man Pun. Test-time intensity consistency adaptation for shadow detection. *arXiv*, 2024.