DeepLIR: Attention-based approach for Mask-Based Lensless Image Reconstruction

Arpan Poudel
University of Arkansas
arpanp@uark.edu

Ukash Nakarmi
University of Arkansas
unakarmi@uark.edu

Abstract

Lensless imaging has emerged as a promising solution to overcome the need for expensive and bulky lenses used in traditional cameras. This technique leverages a mask to optically encode the scene, thus generating a sensor pattern. The image is subsequently reconstructed using a computational algorithm. Traditional model-based reconstruction methods often suffer from prolonged convergence time and subpar perceptual image quality. To mitigate these issues, data-driven deep neural networks can potentially offer enhanced reconstruction quality alongside reduced inference time. However, deep learning methods fall short in providing improved results and tend to produce artifacts, primarily because they do not incorporate any prior knowledge about the imaging model. In this work, we propose a DeepLIR, a hybrid approach that combines the physical system model with a deep learning model. This is achieved by unrolling a conventional model-based optimization algorithm and incorporating an attention-based deep learning model to denoise the image, thereby enhancing the reconstruction quality. Our empirical analysis confirms that DeepLIR surpasses existing lensless image reconstruction techniques in terms of image quality and computational efficiency. Specifically, DeepLIR achieves a remarkable 1.35 × improvement in perceptual quality over the nearest competitor, reflecting its robustness and superiority. Furthermore, it demonstrates superior generalization capabilities when applied to real-world imaging. Code available at : https://github.com/arpanpoudel/lenslessimaging.

1. Introduction

The current imaging system, widely adopted in most modern electronics, is based on a lens that adheres to the pinhole imaging model. However, the physical constraints of the lens hinder its miniaturization, consequently limiting the downsizing of the camera. A potential approach to circumvent this limitation is the utilization of a lensless camera, where a lens is substituted by an optical encoder that captures the scene. The sensor measurements obtained are then reconstructed by computational algorithms. Recent studies have shown promising results in image formation with lensless cameras, which offer smaller size, lower cost, and lighter weight compared to traditional lens-based cameras [3, 5, 18, 26].

Lensless imaging can be categorized into three different systems: illumination-modulated, mask-modulated, and programmable modulator lensless system [6]. In this work, we focus on phase-modulated mask-based lensless imagers where a mask encodes the scene into the sensor measurement and reconstructs the final image with a reconstruction algorithm. Mask-based lensless imaging has applications in 2D imaging [3, 18], 3D imaging [2, 5], and microscopy [1, 9, 19].

The classical approach for image reconstruction involves formulating the imaging system as a model-based inverse problem, which is solved iteratively by minimizing the loss function to reconstruct an image through optimization techniques. This function consists of data fidelity and a regularizer to recover the final image. This approach uses prior information of the optical element, known as the point-spread-function (PSF), which multiplexes the light from the scene. To obtain the PSF of the system, a point light source illuminates the mask and produces a specific pattern on the sensor [25, 27]. However, the multiplexing of light through a mask results in an ill-conditioned system of equations that makes image reconstruction challenging. Iterative methods [4, 7] are used to solve the systems of equations and recover the final image [2]. However, this method doesn't provide a convergence guarantee within a few iterations and produces reconstruction artifacts due to model mismatch, and calibration errors.

The problem of solving the ill-conditioned system of equations for image reconstruction can be modeled using deep learning-based methods. In this approach, a deep neural network (DNN) is utilized to reconstruct an image by learning the parameters of the network, using large datasets of images to map sensor measurements to the ground truth.

However, this method does not consider any prior information about how the images were formed. Additionally, the interpretation of this method can be quite complex. This approach also fails to consider the multiplexing property of the sensor measurement through the mask. Multiplexing is responsible for transforming the local information into overlapping sensor measurements. Consequently, while the deep neural architecture can efficiently identify local features, it does not capture the global features present in the sensor measurement due to the mask's multiplexing property [27] and produce reconstruction artifacts.

In this work, we sidestep this problem by unrolling the fixed number of iterations of the classical algorithm through a neural network and further denoise the resulting image with attention-based U-Net [29] relying on ConvNeXt [22]. We unroll the iterative alternating direction method of multipliers (ADMM) [7] derived with a variable splitting that leverages the specific structure for lensless imaging [2] as a layer of the neural network. The intermediate result is then fed to the denoiser model to reconstruct the final image. Our main idea is to utilize the prior information about the image formation process with the classical approach and use the intermediate information to reconstruct the image with an network with an attention mechanism [32].

We evaluate the performance of our model using images from DiffuserCam [2]. Empirically, we observe that our model outperforms the baselines on lensless images as measured in MSE, PSNR, and LPIPS [37]. Fig. 1 illustrates an overview of our proposed method, and we detail our method further in Section 3; experimental setup and evaluations in Section 4; discussion about advantages and limitations in Section 5; we conclude our work in Section 6.

2. Background And Related Work

2.1. Forward Measurment Model

For a lensless camera, DiffuserCam, we consider the following forward measurement model

$$\begin{aligned} \mathbf{y}(d,x,y) &= \text{crop}[\mathbf{a}(d,x,y) * \mathbf{x}(d,x,y)] \\ &= \mathbf{CAx} \end{aligned} \tag{1}$$

where **a** is a PSF of the imaging system, **x** represents the scene, and (x,y) represents the sensor coordinates for d channels. The symbol * denotes 2D discrete linear convolution and **C** denotes the crop operation to restrict the size of the output. PSF of the DiffuserCam, **a**, is obtained by refracting light from the point source through a diffuser, which creates a high-contrast caustic pattern [2, 18].

The goal is to recover the scene **x** from the measurement **y**. To reconstruct an image **x**, from the measurement **y**, we need to solve Eq. (1) for **x**.

2.2. Unrolling Inverse Algorithm

The reconstructed image can be obtained by solving the following regularized constrained optimization problem:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\geq 0} \frac{1}{2}\|\mathbf{y} - \mathbf{CAx}\|_2^2 + \lambda\|\boldsymbol{\Psi}\mathbf{x}\|_1 \tag{2}$$

Here Ψ denotes the sparsifying transform, such as discrete cosine transform (DCT), finite-differences for total variation (TV), which is a linear operator that transforms the image pixels into sparse representations. λ regularizes the sparsity constraints. A variety of iterative algorithms have demonstrated effectiveness in solving this optimization problem. These include the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [4], the Iterative Shrinkage-Thresholding Algorithm (ISTA), and the Alternating Direction Method of Multipliers (ADMM) [7].

One can construct an ADMM solver with a variable splitting to reconstruct an image from the measurement [2]. To achieve this, Eq. (2) can be formulated as:

$$\begin{aligned} \hat{\mathbf{x}} &= \arg\min_{\mathbf{w}\geq 0, u, v} \frac{1}{2}\|\mathbf{y} - \mathbf{C}v\|_2^2 + \lambda\|u\|_1 \\ &\text{s.t. } v = \mathbf{Ax}, u = \boldsymbol{\Psi}\mathbf{x}, w = \mathbf{x} \end{aligned} \tag{3}$$

With ADMM, the following iteration scheme can be used to reconstruct x^{k+1} at iteration *k*, as formulated in [2],

$$\begin{aligned} u^{k+1} &\leftarrow \Phi_{\frac{\tau}{\mu_2}}\left(\Psi\mathbf{x}^k + \beta^k/\mu_2\right) \\ v^{k+1} &\leftarrow \left(\mathbf{C}^\top\mathbf{C} + \mu_1 I\right)^{-1}\left(\alpha^k + \mu_1\mathbf{A}\mathbf{x}^k + \mathbf{C}^\top\mathbf{y}\right) \\ w^{k+1} &\leftarrow \max\left(\gamma^k/\mu_3 + \mathbf{x}^k, 0\right) \\ \mathbf{x}^{k+1} &\leftarrow \left(\mu_1\mathbf{A}^\top\mathbf{A} + \mu_2\Psi^\top\Psi + \mu_3 I\right)^{-1} r^k \\ \alpha^{k+1} &\leftarrow \alpha^k + \mu_1\left(\mathbf{A}\mathbf{x}^{k+1} - x^{k+1}\right) \\ \beta^{k+1} &\leftarrow \beta^k + \mu_2\left(\Psi\mathbf{x}^{k+1} - u^{k+1}\right) \\ \gamma^{k+1} &\leftarrow \gamma^k + \mu_3\left(\mathbf{x}^{k+1} - w^{k+1}\right), \end{aligned} \tag{4}$$

where

$$\begin{aligned} r^k = &\left(\mu_3 w^{k+1} - \gamma^k\right) + \Psi^\top\left(\mu_2 u^{k+1} - \beta^k\right) \\ &+ \mathbf{A}^\top\left(\mu_1 v^{k+1} - \alpha^k\right). \end{aligned}$$

Here $\Phi_{\frac{\tau}{\mu_2}}$ denotes vectorial soft-thresholding with a threshold value of $\frac{\tau}{\mu_2}$, and $\alpha, \beta,$ and γ represents the Lagrange multipiers associated with *u, v,* and *w*, respectively. μ_1, μ_2 and μ_3 are the penalty parameters.

With this formulation, we incorporate the physical model of the imaging system to reconstruct the image. However, such an approach requires many iterations for convergence, and the resulting images still possess reconstruction artifacts. Furthermore, this method only works well when the priors are correctly chosen. An alternative approach involves executing a predetermined number of iterations, N,

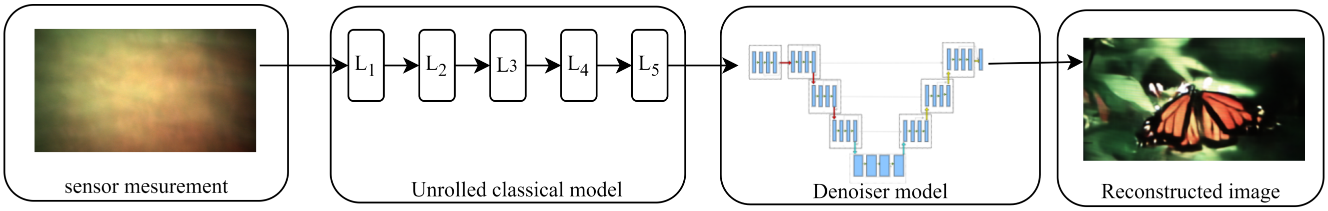

Figure 1. Overview of the reconstruction pipeline with DeepLIR: The process begins with sensor measurements obtained from a lensless camera. The reconstructed image is achieved by solving a constrained optimization problem through a fixed number of iterations. Subsequently, the intermediate result is denoised with a denoiser model.

of these algorithms, with each iteration represented as a layer of a neural network. This can be achieved by making algorithm hyperparameters, such as τ, μ_1, μ_2 and μ_3 as learnable parameters [10] [25]. Many existing works [25] [16] [34] have implemented this idea in lensless image reconstruction. For instance, Le-ADMM [25] has implemented unrolling iterations of ADMM. However, the reconstructed image contains numerous artifacts, is not visually appealing, and scores low on evaluation metrics.

2.3. Denoising models

Image denoising is a challenging computer vision task that aims to remove noise from noisy images. In recent years, the emergence of vision transformers [11] [21], which are based on attention mechanism [32], has shown promising results in solving this task [20].

In contrast to vision transformers, diffusion models [14] [31] with conditional sampling algorithms have demonstrated promising results in solving inverse problems [30] [8] in medical image reconstruction. However, these models are not suitable for lensless imaging systems as they do not incorporate the physical model and often result in reconstruction artifacts. Additionally, the sampling algorithms used in these models can be slow, making them impractical for real-time image generation in lensless cameras.

In the sections that follow, we propose a new image reconstruction pipeline for lensless imaging. This pipeline unrolls the few iterations of ADMM and denoises intermediate images with an attention-based U-Net denoiser model that can bridge the gaps in performance, speed, and compatibility.

3. Method

3.1. Network Architecture

As shown in Fig. 1, Our model combines two models: Unrolled classical model and the Denoiser model.

Unrolled Classical model. Given a sensor measurement $\mathbf{y} \in \mathbb{R}^{H\times W\times C_{in}}$ (H, W and C_{in} are the measurement height, width, and input channel number, respectively), we extract the noisy image $I_N \in \mathbb{R}^{H\times W\times C}$ with a known forward model from y as

$$I_N = H_{UN}(\mathbf{y}) \tag{5}$$

where $H_{UN}(\cdot)$ is the unrolled classical model and it contains N iterations of ADMM represented as a layer in a neural network. More specifically, intermediate images I_1, I_2, ..., I_N are generated after each iteration as

$$I_i = H_{ADMM_i}(I_{i-1}),\ i = 1, 2, ..., N \tag{6}$$

where $H_{ADMM_i}(\cdot)$ denotes the i-th layer of ADMM iteration whose update equation is given by Eq. (4). The trainable parameters for i-th iteration in $H_{ADMM_i}(\cdot)$ are $\mu_1^i, \mu_2^i, \mu_3^i$ and τ^i [25].

Denoiser model. Taking a noisy image I_N, we denoise the image to reconstruct high-quality image I_{RHQ} as

$$I_{RHQ} = H_{DEN}(I_N) \tag{7}$$

where $H_{DEN}(\cdot)$ is the module for reconstruction. Unrolling a few layers of ADMM can be used to iteratively refine the image estimate, while the denoiser model focuses on extracting high-frequency components of an image. To implement the denoiser model, we leverage the attention mechanism [32] on U-Net [29].

Attention-based U-Net: Our denoiser model follows the backbone of DDPM [14] [24] that takes a noisy image and results in a less noisy image. Fig. 2 shows an overview of the model. The following sections provide a thorough discussion of the details of the downsampling, bottleneck, and upsampling operations that are implemented within the U-Net architecture.

3.1.1 Downsampling

The downsampling module within our architecture primarily consists of two ConvNeXt blocks [22], group normalization [33], linear attention [15], a residual connection, and a downsampling operation with a factor of 2. The ConvNeXt block, inspired by research [22], initially employs

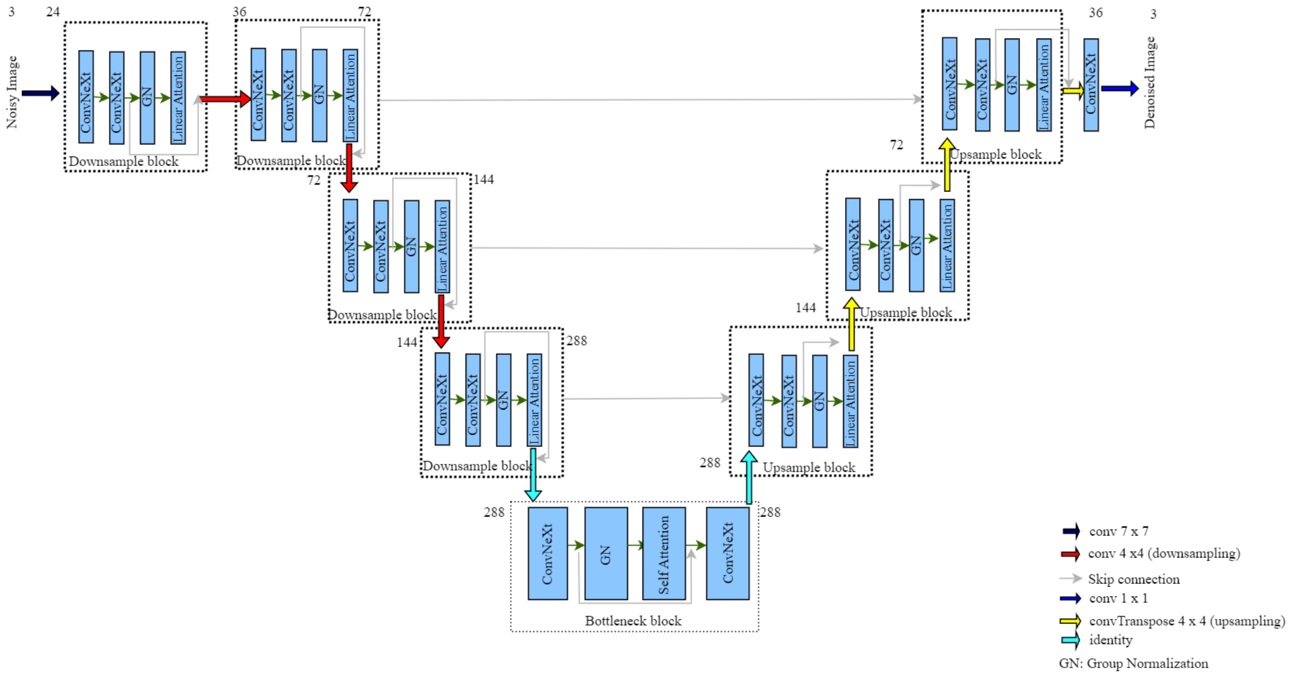

Figure 2. Attention-based U-Net

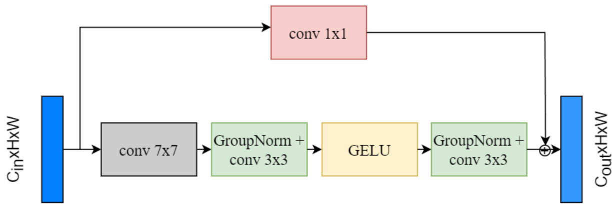

Figure 3. Overview of ConvNeXt block

depthwise separable convolution to effectively capture spatial correlations in the input feature maps. Sequentially arranged convolutional layers, interleaved with GELU activations [13] and group normalization, manipulate these feature maps within this block. The group normalization stabilizes the learning process by mitigating the internal covariate shift [33] within the block. It also incorporates a residual connection, assisting the model in learning identity mappings and preventing the vanishing gradient problem [12]. Fig. 3 gives an overview of ConvNeXt block used in the downsampling module.

The reconstructed images can often be affected by complex distortions or aberrations. These effects may be caused by factors such as the lensless sensor design, environmental conditions, or the properties of the object being imaged. Because of these distortions, important information about an object might be scattered across the entire image. Hence, the need arises for the model to correlate distant regions of the image. This is where the attention mechanism comes in handy. It allows our model to understand the interdependencies between different regions of an image, regardless of their distance in the spatial domain. It essentially provides a form of global context to the model, which can be particularly beneficial for tasks that require an understanding of the whole image, such as denoising with multiplexing property in our case. Hence, the attention mechanism helps capture long-range dependencies in the sensor-acquired noisy image data, and incorporating this layer improves the denoised output, enhancing the clarity and recognizability of key features [11] [15] [36] essential for lensless imaging. To implement an attention mechanism, we use linear attention [15] as it reduces the computational complexity to linear making our model efficient instead of self-attention following the two ConvNeXt and group normalization block.

3.1.2 Bottleneck

The bottleneck module comprises two ConvNeXt blocks with a group normalization and self-attention mechanism and is supported by a residual connection. The self-attention mechanism empowers the model to prioritize the most crucial image features. The residual connection allows the input to bypass the group normalization and self-attention mechanism and be directly added to the output, aiding the model in preserving information from preceding layers and learning identity mappings efficiently.

3.1.3 Upsampling

The upsampling module in our architecture parallels the operations seen in the downsampling module, but with a distinction - the downsampling operation is replaced by an upsampling operation. This change facilitates the reconstruction of the data from the compressed feature space back to its original resolution, aiding in the generation of high-quality outputs.

3.2. Loss function

For DeepLIR, we optimize the parameters of the model by minimizing the mean-squared error (MSE) loss

$$\mathcal{L} = \|I_{HQ} - I_{RHQ}\|_2^2 \tag{8}$$

where I_{HQ} is the ground-truth image and I_{RHQ} is the reconstructed image.

4. Experiments

4.1. Experminetal Setup

Dataset. To train our model, we use the DiffuserCam Lensless Mirflickr Dataset (DLMD) [25] which consists of 25,000 aligned image pairs taken with both DiffuserCam and a lensed camera. We utilize 24,000 image pairs as a training set and 1,000 image pairs as testing images. The raw images are downsampled by a factor of 4 to obtain an image of size 480 × 270 to avoid degradation of lensed image quality due to moiré fringes [25].

We unroll 5 iterations of ADMM with finite differences for total variation (TV) as a sparsifying transform to optimize computation for all our experiments. For attention-based U-Net, we use four layers of the downsampling module, one bottleneck, and three upsampling modules. Following the upsampling module, a final layer which consists of a ConvNeXt and a convolution layer is applied to yield an image of the required shape. This model consists of 19.3 million parameters with 18.3 million trainable parameters.

To compare DeepLIR with other attention-based models, we replaced the denoiser model with SwinIR [20]. After unrolling the classical model, SwinIR is used to refine the reconstructed image. SwinIR key design lies in the use of local-window based self-attention, significantly reducing computation costs and making the network more efficient. For SwinIR, we use a window size of 8, three layers of Residual Swin Transformer Block (RSTB) [20] each containing one Swin Transformer Layer (STL) [21], embedding dimension of 180 and six attention heads. This model consists of 3.1 million parameters with 1.96 million trainable parameters.

Our model is implemented in PyTorch and trained on GeForce RTX 3090 GPU. During the training process, we employed a batch size of 2, and the model was trained for a total of 50 epochs (75 hours to train to completion). We used ADAMW [23] optimizer with the learning rate set to 1×10^{-4}, betas to (0.9, 0.999), and weight decay of 0.01. We used Exponential Moving Average (EMA) on model parameters with a decay factor of 0.995. To perform the quantitative evaluations, we used the 1000 image pairs as testing images, and the inference time was averaged for 100 trials.

4.2. Quantitative Evaluation

In each experiment, we report the Mean Square Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Learned Perceptual Image Patch Similarity (LPIPS) [37] as evaluation metrics. These are utilized to evaluate the quality of the resulting images, comparing them to images obtained with a conventional lensed camera.

Table 1. Comparison of evaluation metrics for various models.

Model	MSE	PSNR	LPIPS	Time (ms)
ADMM (Conv.)	0.15	8.721	0.78	810
Le-ADMM-U [25]	0.0074	21.30	0.1904	75
ADMM+ (SwinIR [20])	0.0069	22.00	0.2069	1140
U-Net [25]	0.0154	18.12	0.2461	10
MMCN [35]	0.0026	25.69	0.1897	91
Kingshott et al. [16]	0.0029	25.34	0.35	84
Rego et al. [28]	0.0087	20.56	-	32
DeepLIR (Ours)	**0.0022**	**26.40**	**0.1412**	165

A comparison of these metrics, as well as inference time on GPU, with those of other models in the field, is presented in Tab. 1, specifically focusing on images captured by the DiffuserCam. In our experimental analysis, we observed that ADMM exhibits prolonged inference times due to its requirement for a larger number of iterations to converge. Despite this extended computational effort, the solutions obtained using ADMM consistently fall behind those achieved by other models in terms of both quality and efficiency. In contrast to ADMM (Converged at 100 iterations) and Le-ADMM-U, DeepLIR produces significantly lower MSE and higher PSNR, indicating more accurate reconstructions. Compared to SwinIR, our method achieves greater performance with a much shorter inference time. While U-Net has the least inference time among the compared methods, its performance in terms of MSE, PSNR, and LPIPS lags significantly behind ours. Lastly, DeepLIR outperforms unrolled primal-dual network [17], model mismatch compensation network [35], and PSF estimation method [28] in all available metrics except for inference time. In the context of lensless imaging, where the quality of visual results holds larger importance, the significance of inference time diminishes compared to the quality of the reconstructed images. While it is true that DeepLIR exhibits a longer inference time compared to competing

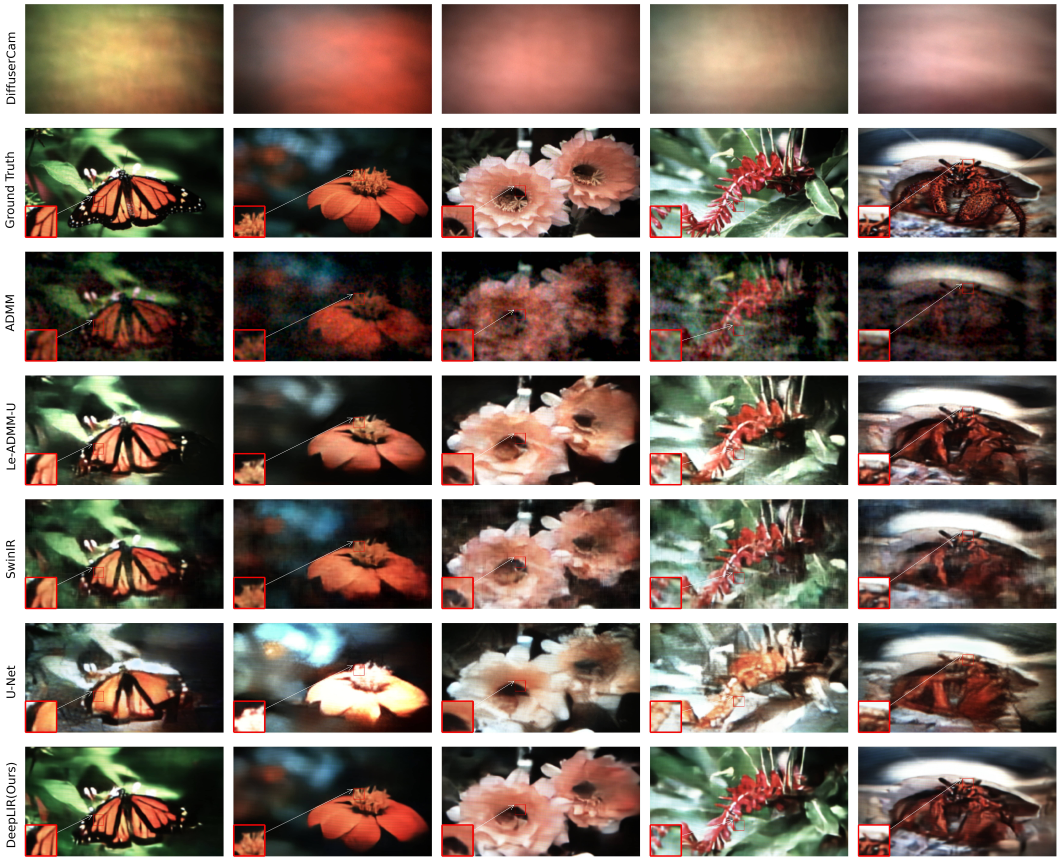

Figure 4. Comparative image reconstructions from DiffuserCam measurements using various models. Big red box: Zoom in version of the indicated smaller red box. DeepLIR is able to reconstruct complex patterns and fine details where other models fail. You may zoom in to view more details.

methods like unrolled primal-dual networks [17], model mismatch compensation networks [35], and PSF estimation methods [28], its superior performance across all other available metrics underscores its effectiveness. In summary, DeepLIR achieves the best MSE, LPIPS, and PSNR with an inference time of 165 ms which demonstrates a satisfactory balance between performance and computational efficiency.

4.3. Qualitative Evaluation

In addition to the quantitative analysis, we conduct a qualitative evaluation to visually examine the quality of image reconstruction by our model compared to the other methods. The comparison is illustrated in Fig. 4.

The reconstructed images from our DeepLIR model visibly outperform those from other methods. Specifically, the images produced by our model exhibit superior clarity, sharper details, and improved color fidelity. This is particularly evident when observing intricate details and color gradients in the reconstructed images. In comparison, the ADMM and Le-ADMM-U models appear to introduce more noise and distortion into the images. The reconstructions by SwinIR and U-Net also present issues with clarity, and in the case of U-Net, there is a noticeable loss of image detail and color accuracy. The visual results demonstrate that our DeepLIR model not only excels in quantitative metrics but also delivers superior visual quality in image reconstruction.

4.4. Generalization to real-world imaging

Following our qualitative evaluations, we test DeepLIR's adaptability to real-world scenarios. The primary challenge in lensless imaging is dealing with the unpredictabilities of natural environments—variations in ambient light, dynamic subjects, and obstructions that introduce noise. We provide a visualization of the real-world image reconstruction Fig. 5. Remarkably, DeepLIR showcased a good performance when deployed in these uncontrolled settings. It effectively managed to reconstruct subjects and details from raw diffraction patterns, even in conditions far from the ideal parameters of our training data and test set taken from a monitor screen.

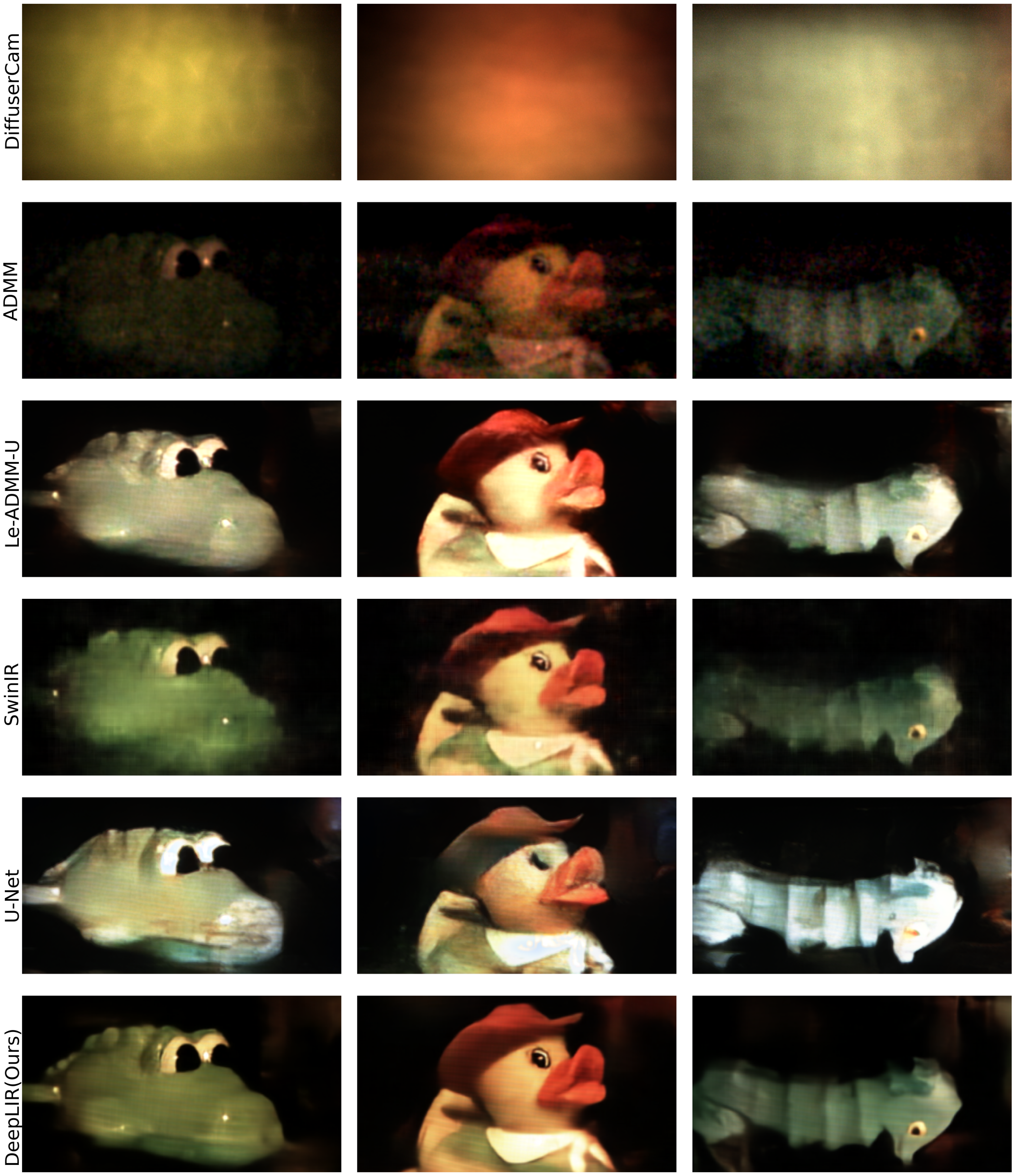

Figure 5. Comparative reconstructions of real-world scenes from DiffuserCam measurements using various models. Among all the tested methods, only DeepLIR manages to reconstruct the images with remarkably accurate colorization and fewer artifacts. For example, the shadow of the hat in the second image appears more realistic with DeepLIR. The texture of the alligator's skin in the first image is also better represented with DeepLIR. Additionally, the seahorse in the third image is best depicted with DeepLIR, while other models struggle with this particular detail.

5. Discussion

The presented experiments and results have led us to several key observations about the DeepLIR model, its performance, and its relation to other existing methods in the field of image reconstruction from DiffuserCam measurements.

5.1. Advantages of DeepLIR

Balanced Performance: One of the most significant takeaways from our quantitative evaluations is that DeepLIR offers a robust balance between image quality and computational speed. While there are methods with shorter inference times, such as Le-ADMM-U, they lag behind significantly in terms of image quality. On the other hand, while

SwinIR provides good image quality, its long inference time makes it less suitable for real-time applications.

Visual Quality: The qualitative evaluations emphasize the high visual quality of reconstructions achieved by DeepLIR. Not only does it offer sharper details and better color fidelity, but it also outperforms other methods when faced with intricate details, which are often a challenge for lensless imaging.

5.2. Possible Improvements and Extensions

Unsupervised approach: In our current method, the measurements from the DiffuserCam heavily rely on the specific PSF, which encapsulates the physical model of our measurement procedure. A major limitation arises when altering the measurement process, such as by swapping to a different mask. Such alterations necessitate the collection of new paired images in alignment with the modified process, followed by a comprehensive retraining of our DeepLIR model. This inherent limitation restricts DeepLIR's adaptability to diverse measurement processes.

To overcome this challenge, our future endeavors will shift towards an unsupervised approach. Our plan is to harness the power of generative models to learn the inherent distribution of real-world images. By achieving this, we aim to provide an efficient sampling algorithm capable of reconstructing images in real time, regardless of the specific measurement process in play.

Real-world imaging: In our experiments with DiffuserCam's measurement, while results were promising in controlled settings, challenges arose in real-world imaging scenarios, particularly in the precise reconstruction of background details. Several factors might be at play, including the inherent complexity of real-world scenes, the current lensless imaging limitations, and external noise. Future work will aim to address these by considering adaptive PSFs tailored to general scene conditions, integrating more discerning algorithms for detailed reconstructions, and employing advanced noise-reduction techniques. Additionally, enhancements in DiffuserCam's optical components and sensor sensitivity could pave the way for better real-world imaging fidelity.

6. Conclusion

In conclusion, we introduce a new pipeline to address lensless image reconstruction, unrolling several iterations of ADMM algorithms to procure an initial estimate of the scene. Leveraging this estimate, an attention-based U-Net is applied for denoising, wherein the attention mechanism assists in retrieving information dispersed by the inherent multiplexing property of light. Empirical evaluations exhibit that DeepLIR provides reconstructions of superior perceptual quality across both controlled environments and real-world imaging scenarios. In contrast to existing work, DeepLIR offers noticeable advantages in both qualitative and practical aspects. The findings of this work shed light on the promising synergy between the attention mechanism and classical methods, highlighting their potential to enhance performance beyond that of larger, previously successful models in image reconstruction and denoising. This research not only advances the state-of-the-art in lensless image reconstruction but also underscores the broad applicability and efficiency of integrating attention mechanisms within traditional frameworks.

References

[1] Jesse K. Adams, Vivek Boominathan, Benjamin W. Avants, Daniel G. Vercosa, Fan Ye, Richard G. Baraniuk, Jacob T. Robinson, and Ashok Veeraraghavan. Single-frame 3d fluorescence microscopy with ultraminiature lensless flatscope. *Science Advances*, 3(12):e1701548, 2017. 1

[2] 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, Jan 2018. 1, 2

[3] 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. 1

[4] 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. 1, 2

[5] 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. 1

[6] Vivek Boominathan, Jacob T. Robinson, Laura Waller, and Ashok Veeraraghavan. Recent advances in lensless imaging. *Optica*, 9(1):1–16, Jan 2022. 1

[7] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. 2011. 1, 2

[8] Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated mri, 2022. 3

[9] Ashwin Dhakal, Rajan Gyawali, Liguo Wang, and Jianlin Cheng. Cryoppp: A large expert-labelled cryo-em image dataset for machine learning protein particle picking. *bioRxiv*, 2023. 1

[10] Steven Diamond, Vincent Sitzmann, Felix Heide, and Gordon Wetzstein. Unrolled optimization with deep priors. *CoRR*, abs/1705.08041, 2017. 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 3, 4

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 4

[13] Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415, 2016. 4

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *CoRR*, abs/2006.11239, 2020. 3

[15] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. *CoRR*, abs/2006.16236, 2020. 3, 4

[16] Oliver Kingshott, Nick Antipa, Emrah Bostan, and Kaan Akşit. Unrolled primal-dual networks for lensless cameras. *Opt. Express*, 30(26):46324–46335, Dec 2022. 3, 5

[17] Oliver Kingshott, Nick Antipa, Emrah Bostan, and Kaan Akşit. Unrolled primal-dual networks for lensless cameras. *Opt. Express*, 30(26):46324–46335, Dec 2022. 5, 6

[18] Grace Kuo, Nick Antipa, Ren Ng, and Laura Waller. Diffusercam: Diffuser-based lensless cameras. In *Imaging and Applied Optics 2017 (3D, AIO, COSI, IS, MATH, pcAOP)*, page CTu3B.2. Optica Publishing Group, 2017. 1, 2

[19] Grace Kuo, Fanglin Linda Liu, Irene Grossrubatscher, Ren Ng, and Laura Waller. On-chip fluorescence microscopy with a random microlens diffuser. *Opt. Express*, 28(6):8384–8399, Mar 2020. 1

[20] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer, 2021. 3, 5

[21] 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. *CoRR*, abs/2103.14030, 2021. 3, 5

[22] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *CoRR*, abs/2201.03545, 2022. 2, 3

[23] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. 5

[24] Lucidrains. Implementation of denoising diffusion probabilistic model in pytorch. https://github.com/lucidrains/denoising-diffusion-pytorch, 2020. GitHub repository. 3

[25] Kristina Monakhova, Joshua Yurtsever, Grace Kuo, Nick Antipa, Kyrollos Yanny, and Laura Waller. Learned reconstructions for practical mask-based lensless imaging. *Opt. Express*, 27(20):28075–28090, Sep 2019. 1, 3, 5

[26] Aydogan Ozcan and Euan McLeod. Lensless imaging and sensing. *Annual Review of Biomedical Engineering*, 18(1):77–102, 2016. PMID: 27420569. 1

[27] Xiuxi Pan, Xiao Chen, Saori Takeyama, and Masahiro Yamaguchi. Image reconstruction with transformer for mask-based lensless imaging. *Opt. Lett.*, 47(7):1843–1846, Apr 2022. 1, 2

[28] Joshua D. Rego, Karthik Kulkarni, and Suren Jayasuriya. Robust lensless image reconstruction via psf estimation. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 403–412, 2021. 5, 6

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. 2, 3

[30] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models, 2022. 3

[31] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *CoRR*, abs/2011.13456, 2020. 3

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. 2, 3

[33] Yuxin Wu and Kaiming He. Group normalization. *CoRR*, abs/1803.08494, 2018. 3, 4

[34] Jingyu Yang, Xiangjun Yin, Mengxi Zhang, Huihui Yue, Xingyu Cui, and Huanjing Yue. Learning image formation and regularization in unrolling amp for lensless image reconstruction. *IEEE Transactions on Computational Imaging*, 8:479–489, 2022. 3

[35] 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. 5, 6

[36] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks, 2019. 4

[37] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *CoRR*, abs/1801.03924, 2018. 2, 5