FourierNets enable the design of highly non-local optical encoders for computational imaging

Anonymous Author(s)
Affiliation
Address
email

Abstract

Differentiable simulations of optical systems can be combined with deep learning-based reconstruction networks to enable high performance computational imaging via end-to-end (E2E) optimization of both the optical encoder and the deep decoder. This has enabled imaging applications such as 3D localization microscopy, depth estimation, and lensless photography via the optimization of local optical encoders. More challenging computational imaging applications, such as 3D snapshot microscopy which compresses 3D volumes into single 2D images, require a highly non-local optical encoder. We show that existing deep network decoders have a locality bias which prevents the optimization of such highly non-local optical encoders. We address this with a decoder based on a shallow neural network architecture using global kernel Fourier convolutional neural networks (FourierNets). We show that FourierNets surpass existing deep network based decoders at reconstructing photographs captured by the highly non-local DiffuserCam optical encoder. Further, we show that FourierNets enable E2E optimization of highly non-local optical encoders for 3D snapshot microscopy. By combining FourierNets with a large-scale multi-GPU differentiable optical simulation, we are able to optimize non-local optical encoders 170× to 7372× larger than prior state of the art, and demonstrate the potential for ROI-type specific optical encoding with a programmable microscope.

1 Introduction

Modern computational optics relies on the end-to-end (E2E) optimization of the optical system (optical encoder) and the computational image reconstruction algorithm (computational decoder). This has been enabled by the development of differentiable physics-based simulations of optical encoders, paired with deep network based decoders. This E2E approach has been successfully applied to the estimation of depth from defocus [1, 2, 3, 4], lensless photography [5], and particle localization microscopy [6, 7]. Despite this wide success, only the space of local optical encoders has been explored in this fashion due to the computational expense of physics-based simulation of highly non-local optical encoders [3]. This is particularly the case for 3D snapshot microscopy which is well-known to require highly non-local optical encoders, and for which no E2E solutions have yet been demonstrated.

We show that there are actually two issues blocking the E2E optimization of highly non-local optical encoders. First, the computational expense of physics-based simulation of non-local optical encoders is significantly greater, since they are orders of magnitude larger than local encoders. And second, the locality bias inherent to deep network architectures currently used for computational imaging prevents decoding from non-local solutions. In this paper, we primarily focus on developing a computational framework for engineering highly non-local optical encoders by solving both issues. We solve the

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

first issue by developing a large-scale multi-GPU differentiable optical simulation, and the second by developing a new FourierNet architecture based deep network decoder.

In 3D snapshot microscopy, all existing methods [8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 5, 19] employ a highly non-local optical encoding to optically transform a 3D volume into a single 2D camera image, which is computationally decoded to reconstruct the 3D volume. Fast volumetric imaging is invaluable across biology, including for imaging neural activity across the whole brain of an animal [20, 21, 22]. Here, we focus on developing a 3D snapshot microscope for imaging neuronal activity across the whole brain of the larval zebrafish (*Danio rerio*) at camera rates exceeding 100Hz. This is two orders of magnitude faster than the fastest conventional microscopes — light sheet imaging of whole brain neural activity can only achieve 0.5Hz - 2Hz volume rates [20]. This improvement of temporal resolution is essential for imaging with fast calcium indicators [23] and voltage indicators [24].

Newly developed programmable microscopes with up to 10^6 free parameters, e.g. pixels on a spatial light modulator (SLM), enable the implementation of a rich space of optical encodings, and present the possibility of direct optimization of snapshot microscope parameters specifically for particular ROI types and imaging tasks. While our paper primarily focuses on addressing the computational challenges of E2E optimization of highly non-local optical encoders, we also demonstrate proof-of-concept that an SLM-based programmable microscope can implement such an engineered non-local optical encoder, paving the way to implementing ROI-type and task specific optical encoders in a single physical microscope. In contrast, 3D snapshot imaging has classically been performed using fixed optical encoders based on microlens arrays [8, 9, 10, 11, 12, 13, 14, 15], pseudorandom diffusers/masks [16, 17], and designed or optimized diffusers/masks [18, 5, 19], with some of these methods using deep learning for reconstruction [11, 15].

Problem statement We define an optical encoder as $\mathbf{M}_{\boldsymbol{\phi}}$ parameterized by $\boldsymbol{\phi}$ and a computational decoder as $\mathbf{R}_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$. Optical encoders can usually be modeled as linear transfer functions implemented by the optics. In special cases, such as for the optical systems explored in this paper, they can be represented as a linear convolution filter called a "point spread function" which has its coefficients computed by a wave optics simulation, dependent on the physical parameters $\boldsymbol{\phi}$ of the optical system (Appendix A.1). We wish to develop a reconstruction network architecture for decoding non-local encodings in 2D images $\mathbf{c}$ produced by $\mathbf{M}_{\boldsymbol{\phi}}$ and also to enable the end-to-end optimization of $\boldsymbol{\phi}$ to produce non-local encodings. We'll consider two applications to investigate these optimization problems: (1) lensless photography (Figure 2), where we optimize only the reconstruction network $\mathbf{R}_{\boldsymbol{\theta}}$ to reconstruct images of natural scenes from camera images captured by the DiffuserCam [25], and (2) 3D snapshot microscopy using an SLM-based programmable microscope (Figure 1A), where we perform E2E optimization of both the reconstruction network $\mathbf{R}_{\boldsymbol{\theta}}$ and the optical encoding $\mathbf{M}_{\boldsymbol{\phi}}$ based on the SLM parameters in order to image 3D volumes $\mathbf{v}$ and reconstruct 3D volumes $\hat{\mathbf{v}}$. While we focus on solving the computational challenges involved in the E2E optimization of large highly non-local optical encoders, we also demonstrate a proof-of-concept implementation of our designed optical encoder using a spatial light modulator based programmable microscope.

1.1 Our contributions

1. We developed a large-scale, parallel, multi-GPU differentiable wave optics simulation of a programmable microscope, based on a $4f$ optical model with a phase mask ($\boldsymbol{\phi}$) implemented using a spatial light modulator (SLM), described further in Appendix A.1. SLMs $\boldsymbol{\phi}$ can have over 10^6 optimizable parameters, which we can feasibly simulate and optimize to produce PSFs with $170\times$ to $7372\times$ more unique voxels than previous attempts at deep learning PSF optimization using single GPUs [6, 3] (Appendix A.5).

2. We collected a large dataset of high resolution 3D confocal volumes of zebrafish larvae for the purpose of ROI-type specific end-to-end optimization of optical encoders.

3. We introduce an efficient FourierNet reconstruction network architecture for decoding from non-local optical encoders using very large global convolutions implemented via Fourier convolutions.

4. We show that our networks outperform the state-of-the-art deep decoders for DiffuserCam based lensless photography [25] and for 3D snapshot microscopy.

5. Our method enables, for the first time, direct end-to-end optimization of highly non-local optical encoders in the space of spatial light modulator (SLM) pixels with over 10^6 parameters. In simulation, we demonstrate the potential for significant improvements in imaging resulting from ROI-type specific optimization of optical encoders.

1.2 Prior work

Neural network architectures for computational imaging have all used convolution layers with small filters. End-to-end optimization of optical encoders have largely been performed with UNet-based architectures [1, 2, 3, 4, 5, 7], with one method using a ResNet-based architecture [6]. Such optimization has always led to local optical encodings. End-to-end optimization has never been attempted for large-field of view 3D snapshot microscopy due to the difficulty of simulating and reconstructing from non-local encoders. However, small filter convolutional deep networks have been used in a non-end-to-end manner to reconstruct volumes from 3D snapshot microscopes designed using microlens arrays [15, 26, 27]. One recent hybrid approach combines a UNet with a differentiable approximate inverse method (Wiener filter) to handle non-local spatially varying (non-convolutional) optical encoders [27]. For photography and MRI, another hybrid approach of deep learning combined with unrolling iterations of traditional deconvolution algorithms provides the benefits of fast amortized optimization and higher quality reconstructions due to learning of structural priors [28, 29, 25, 26]. We note that these hybrid approaches typically require measurement of the optical encoder of the system, whereas our method does not and can learn to produce high quality reconstructions using only pairs of ground truth and system images.

In our work, we demonstrate that convolution layers with large filters implemented efficiently in the Fourier domain enable the end-to-end learning of highly non-local optical encoders for 3D snapshot microscopy. Large convolution filters have been shown to be helpful for other computer vision applications such as semantic segmentation and salient object detection [30, 31], and the Fourier domain parameterization of small filters has been described previously [32].

A pioneering strategy in 3D snapshot microscopy has been light field microscopy [8], which employs a microlens array at the microscope's image plane to create subimages encoding both amplitude and phase of light [8, 33]. A variety of microlens-array-based light field microscopes have been used to perform whole-brain imaging [8, 11, 14, 22, 13, 12, 9, 34, 19]. [19] optimizes the placement of microlenses, but not in an end-to-end manner. Despite variation in design, microlens-based microscopes have, to various degrees, four main limitations that can be improved: 1) blocking or scattering of light between microlenses, causing light inefficiency, 2) not making use of all pixels on the camera to encode a 3D volume, leading to inefficient compression and suboptimal reconstructions $\hat{\mathbf{v}}$, 3) aliasing at some planes with generally nonuniform axial resolution, and 4) a fixed optical encoding scheme.

An alternative to using microlenses is to implement a coded detection strategy using a phase mask or diffuser to spread light broadly across the camera sensor [16, 18, 35, 5, 10, 17, 19]. The designed phase masks can be implemented either by manufacturing a custom optical element or using a programmable SLM [2, 3, 4, 5, 19, 6]. Using a programmable element allows different microscope parameters to be used for different sample and ROI types.

2 Methods

We show our network architecture and an overview of autoencoder training both the microscope parameters $\boldsymbol{\phi}$ and reconstruction network parameters $\boldsymbol{\theta}$ in Figure 1A. The programmable microscope is simulated by a differentiable implementation of a wave-optics model of light propagation. We have selected a programmable microscope design based on pupil-plane phase modulation with a programmable spatial light modulator, for which imaging is well-approximated by a computationally-efficient convolution [36]. A detailed description of our simulation is provided in Appendix A.1. For lensless photography, there is no optical simulation because the images have been collected on a real camera.

2.1 FourierNet for decoding non-local optical encoders

Because images created by optical encoders can potentially encode signals from the incoming light field to any location in the camera image in a non-local manner, it is essential that a reconstruction network have global context. Existing multi-scale architectures such as the UNet [37] can achieve

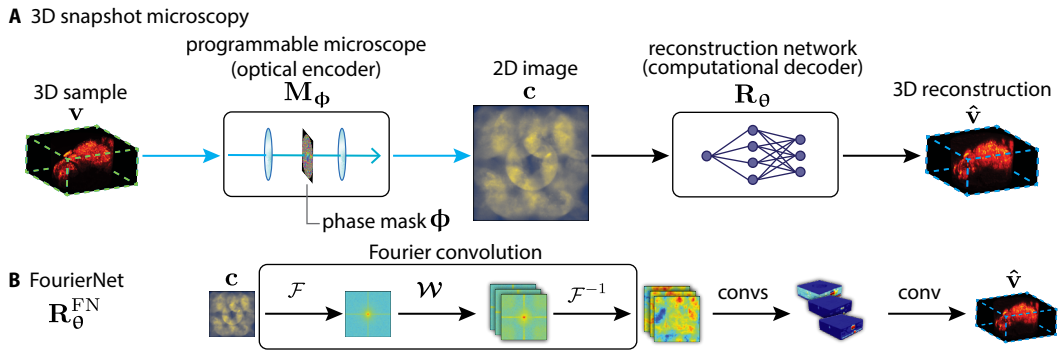


Figure 1: Overview of our problem setup and our proposed network architectures. Top row (**A**) shows the problem of 3D snapshot microscopy, where we computationally reconstruct a 3D volume from a 2D image. Bottom row (**B**) shows our proposed FourierNet architecture, which includes a **Fourier convolution** layer that enables efficient computation of global features.

global context, but at the expense of many computation layers with small convolution kernels which we speculate to have a local information bias which is inappropriate for computational optics. In this paper, we introduce relatively shallow architectures for computational imaging, which rely on convolution layers with very large filters, implemented efficiently in the Fourier domain [32].

FourierNet We propose a simple three layer convolutional network architecture with very large global convolutions at the very first layer, followed by two standard local convolutional layers (Figure 1B). We define a global convolution as a convolution with kernel size equal to the input image. Such a convolution achieves global context in a single step but is computationally expensive. Global convolutions are implemented more efficiently in the Fourier domain, yielding a speed up of two orders of magnitude. Due to the use of Fourier convolutions to enable global context, we call our architecture FourierNet. In contrast to a typical UNet which can contain many tens of convolution layers, the FourierNet is only three layers deep, which requires backpropagation through fewer layers compared to a typical UNet with the same receptive field.

Fourier domain convolutions It is well-known that large kernel convolutions can be implemented more efficiently in the Fourier domain [38, 39, 40]. A naive implementation of global convolution requires $\mathcal{O}(N^2)$ operations, where N is number of pixels in both the image and the kernel. An alternative global convolution implementation is to Fourier transform $\mathcal{F}$ the input $\mathbf{x}$ and convolution kernel $\mathbf{w}$, perform element-wise multiplication in Fourier space, and finally inverse Fourier transform $\mathcal{F}^{-1}$, requiring only $\mathcal{O}(N \log N)$ operations [39, 40]. Following [32], we store and optimize the weights in Fourier space $\boldsymbol{\mathcal{W}}$. This over-parameterization costs $8\times$ the memory of an equivalent real valued large filter but saves the computational cost of Fourier transforming the real-valued weights (Appendix A.6). Thus a Fourier convolution is defined:

$$\mathbf{Re}\{\mathcal{F}^{-1}\{\boldsymbol{\mathcal{W}} \odot \mathcal{F}\{\mathbf{x}\}\}\} \tag{1}$$

For image and kernel sizes of 256×256 pixels, our implementation leads to nearly $500\times$ speedup: standard PyTorch convolution takes 2860ms, while Fourier convolution takes 5.92ms on a TITAN X. We also show how we can naturally extend our Fourier convolutions and FourierNet to a multiscale version using cropping in the Fourier domain in Appendix A.2.

2.2 Physics-based autoencoder for simultaneous engineering of microscope encoder and reconstruction network decoder

We describe the imaging process as the following transformation from the 3D light intensity volume $\mathbf{v}$ to the 2D image formed on the camera $\mathbf{c}$:

$$\boldsymbol{\mu}_\mathbf{c} = \mathbf{M}_\boldsymbol{\phi}(\mathbf{v}) \tag{2}$$

$$\mathbf{c} = \max\left([\boldsymbol{\mu}_\mathbf{c} + \sqrt{\boldsymbol{\mu}_\mathbf{c}}\epsilon], 0\right), \epsilon \sim \mathcal{N}(0, 1) \tag{3}$$

where $\mathbf{M}_\boldsymbol{\phi}$ denotes the microscope parameterized by a 2D phase mask, $\boldsymbol{\phi}$. This phase mask $\boldsymbol{\phi}$ describes the 3D-to-2D encoding of this microscope model completely. A Poisson distribution with mean rate $\boldsymbol{\mu}_\mathbf{c}$ describes the physics of photon detection at the camera, but sampling from this

distribution is not differentiable. We approximate the noise distribution with a rectified Gaussian. We include details on $\mathbf{M}_{\mathbf{\Phi}}$ in Appendix A.1 [36]. Jointly training reconstruction networks and microscope parameters involves image simulation, reconstruction, then gradient backpropagation to update the reconstruction network and microscope parameters. Our parallelization strategy enables optimization of phase masks with millions of parameters in a feasible amount of time and memory per GPU. This also enables us to produce PSFs with multiple orders of magnitude more unique voxels than previous attempts (Appendix A.4). Details on parallelization, planewise reconstruction networks, and planewise sparse gradients are provided in Appendix A.4, Figure 12.

We use the the normalized mean squared error (NMSE) as the basis for our loss function. Since we care more about the high spatial frequency content of the image which contains mostly cells, and less about the low spatial frequencies which contain mostly background, we implemented a two part loss function. L_{HNMSE} measures the NMSE between the high pass filtered volume and high pass filtered reconstruction. This is combined with the unfiltered NMSE L_{NMSE} between the original volume and its reconstruction. Formally, our loss function $L(\mathbf{v}, \hat{\mathbf{v}})$ for all snapshot microscopy reconstruction problems is defined as:

$$L(\mathbf{v}, \hat{\mathbf{v}}) = L_{\text{HNMSE}}(\mathbf{v}, \hat{\mathbf{v}}) + \beta L_{\text{NMSE}}(\mathbf{v}, \hat{\mathbf{v}}) \tag{4}$$

$$L_{\text{HNMSE}}(\mathbf{v}, \hat{\mathbf{v}}) = \frac{\mathbb{E}\left[(H(\mathbf{v}) - H(\hat{\mathbf{v}}))^2\right]}{\mathbb{E}(H(\mathbf{v})^2)}, L_{\text{NMSE}}(\mathbf{v}, \hat{\mathbf{v}}) = \frac{\mathbb{E}\left[(\mathbf{v} - \hat{\mathbf{v}})^2\right]}{\mathbb{E}(\mathbf{v}^2)} \tag{5}$$

where $H(\cdot)$ denotes high pass filtering and $\mathbb{E}(\cdot)$ denotes the mean over pixels and sample volumes. Both loss terms are normalized as shown to reduce variance in L, which can otherwise cause large magnitude fluctuations based on the brightness variation across training volumes and cause training instability. For our experiments, we set the weight β for the L_{NMSE} term to 0.1.

3 Results

DiffuserCam Lensless Mirflickr Dataset We also test reconstruction performance on experimental computational photography data[1] from [25] (Figure 2). This is a dataset constructed by displaying RGB color natural images from the MIRFlickr dataset on a monitor and then capturing diffused images by the DiffuserCam lensless camera. The dataset contains 24,000 pairs of DiffuserCam and ground truth images. The goal of the dataset is to learn to reconstruct the ground truth images from the diffused images. As in [25], we train on 23,000 paired diffused and ground truth images, and test on 999 held-out pairs of images.

Larval Zebrafish Snapshot Microscopy Dataset We show all our results for snapshot microscopy using our simulation of a snapshot microscope, and our engineered optical encoders have not been experimentally tested on a programmable microscope. We simulate snapshot imaging using high resolution confocal imaging volumes of zebrafish. These are volumes of transgenic larval zebrafish whole brains expressing nuclear-restricted GCaMP6 calcium indicator in all neurons. These images are representative of brain-wide activity imaging. We train on 58 different zebrafish volumes (which we augment heavily) and test on 10 held-out volumes. For all experiments, we downsample the high resolution confocal data to (1.0 µm z, 1.625 µm y, 1.625 µm x). We created 4 datasets from these scanned volumes corresponding to imaging different fields of view or regions of interest for our experiments. Full specifications for these datasets are in Table 6 (Appendix A.4). For Figure 3, we restrict the field of view to (200 µm z, 416 µm y, 416 µm x) with a tall cylinder cutout of diameter 193 µm and height 200 µm and image with 256×256 pixels on the simulated camera sensor. Figure 4 and Table 3 show our larger experiments with 512×512 simulated camera pixels, with a field of view of (250 µm z, 832 µm y, 832 µm x).

3.1 FourierNets outperform state-of-the-art for reconstructing natural images captured by DiffuserCam lensless camera

We compare our FourierNet architecture to the best learned method from [25] using an unrolled ADMM and a denoising UNet, as well as to a vanilla UNet from [25]. Architecture details are in Appendix A.6, A.7. We can see that FourierNet visually outperforms the methods from [25] in Figure 2, and quantitatively in Table 1 across all quality metrics. The Le-ADMM-U and UNet results in Table 1 were reported by [25] using a combined MSE + LPIPS loss. Unlike [25], we find that training

[1] Publicly available: https://waller-lab.github.io/LenslessLearning/dataset.html

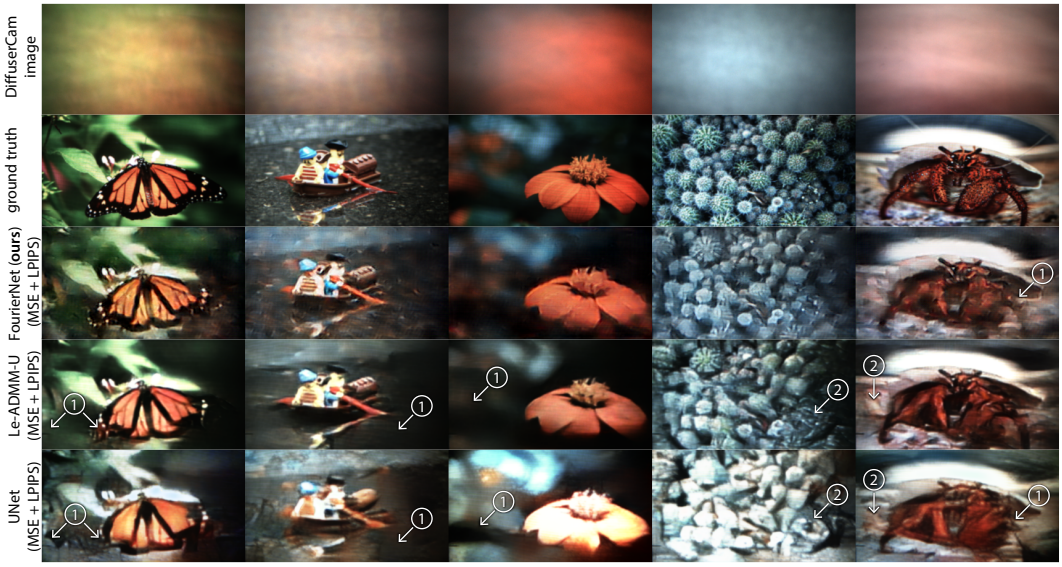


Figure 2: Comparisons of our method (third row) to state-of-the-art learned reconstruction methods on lensless diffused images of natural scenes. Regions labeled (1) indicate missing details, either resolution or textures in backgrounds. Regions labeled (2) indicate hallucinated textures. Note that the previous state-of-the-art solutions [25] exhibit both issues more often compared to our models.

Table 1: Quality of natural image reconstruction on the DiffuserCam Lensless Mirflickr Dataset (mean $\pm$ s.e.m., $n = 999$). Superscripts denote loss function: [1] MSE, [2] MSE+LPIPS.

Method	**MSE** $\downarrow$ ($\times 10^{-2}$)	**LPIPS** $\downarrow$	**MS-SSIM** $\uparrow$	**PSNR** $\uparrow$	**Time** $\downarrow$ (ms)
FourierNet[1]	**0.39** $\pm$ **0.007**	0.20 $\pm$ 0.00	**0.882** $\pm$ **0.001**	**24.8** $\pm$ **0.09**	35.54
FourierNet[2]	0.54 $\pm$ 0.010	**0.16** $\pm$ **0.00**	0.868 $\pm$ 0.001	23.4 $\pm$ 0.09	35.54
Le-ADMM-U[2] [25]	0.75 $\pm$ 0.021	0.19 $\pm$ 0.00	0.865 $\pm$ 0.002	22.1 $\pm$ 0.09	48.59
UNet[2] [25]	1.68 $\pm$ 0.060	0.24 $\pm$ 0.00	0.818 $\pm$ 0.002	19.2 $\pm$ 0.11	**06.97**

FourierNets with the MSE loss alone provides reconstructions visually similar to the ground truth as shown in Figure 13 (Appendix A.7). Timings in Table 1 are for only the forward pass on a single TITAN Xp GPU.

3.2 FourierNets outperform UNets for engineering non-local optical encoders

Table 2: Quality of reconstructed volumes after optimizing microscope parameters to image zebrafish on 256×256 pixel camera (mean $\pm$ s.e.m., $n = 10$)

Microscope	**Reconstruction**	L_{HNMSE} $\downarrow$	**MS-SSIM** $\uparrow$	**PSNR** $\uparrow$	**Time** $\downarrow$ (s)
FourierNet2D	FourierNet3D	**0.6093** $\pm$ **0.0209**	**0.955** $\pm$ **0.004**	**34.89** $\pm$ **0.88**	**0.38**
FourierNet2D	UNet3D	0.7298 $\pm$ 0.0151	0.923 $\pm$ 0.008	30.16 $\pm$ 0.94	0.96
Wiener + UNet	Wiener + UNet	0.7223 $\pm$ 0.0179	**0.957** $\pm$ **0.003**	**34.49** $\pm$ **0.91**	0.73
UNet2D	UNet3D	0.7109 $\pm$ 0.0161	0.913 $\pm$ 0.009	29.17 $\pm$ 1.13	0.96

We compare optimizing microscope parameters $\boldsymbol{\phi}$ with three neural networks: 1) using our FourierNet with 2D convolutions (FourierNet2D), 2) using a vanilla UNet with 2D convolutions (UNet2D), and 3) using a Wiener deconvolution as an approximate inverse combined with a refining UNet, as in [3]. Training is a two stage process in which a plane-wise reconstruction network with fewer parameters and 2D convolutions is used during optimization of $\boldsymbol{\phi}$, then once $\boldsymbol{\phi}$ is fixed a more powerful reconstruction network with 3D convolutions is used, except for the Wiener + UNet [3]

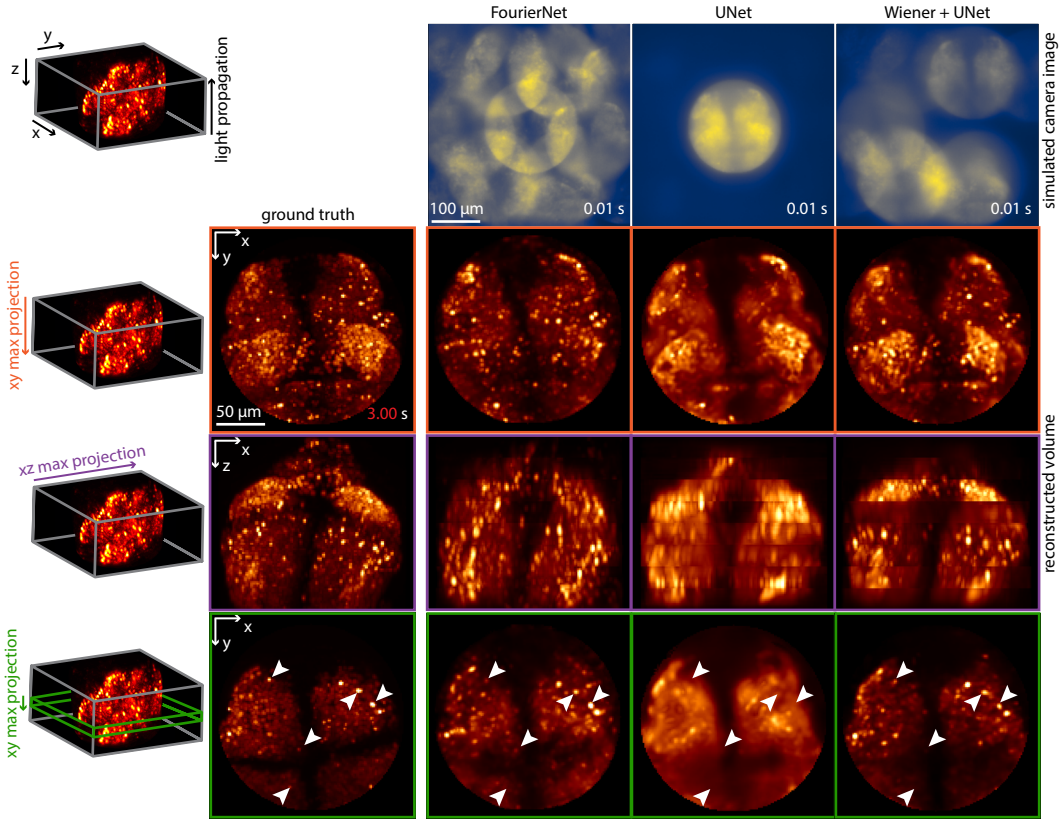
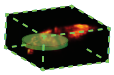
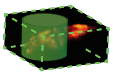
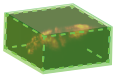


Figure 3: Comparing simulated camera images (0.01 second expected acquisition time) and corresponding reconstructions of a volume captured using our FourierNet (left) versus UNet (middle) and Wiener + UNet [3] (right) optimized microscopes. Top row shows simulated 256×256 pixel camera images; bottom right of camera image shows approximate acquisition time (given a reasonable number of simulated photons per pixel, i.e. SNR). Ground truth has no corresponding camera image, because the 3D volume is imaged directly via slow high resolution confocal microscopy (3 second acquisition time). Colored arrows in left column show projection axis for each row. White arrows show individual neurons clearly visible for FourierNet, but not for UNet. Wiener + UNet [3] has detail in some planes, but not consistently throughout the whole volume, and also under-utilizes camera pixels.

method which is trained in a single stage as in [3]. Table 2 shows the type of network used for the first stage to train the optical encoder $\mathbf{M}_{\boldsymbol{\Phi}}$ in the first column, and the type of network used in the second stage for training the reconstruction network only (given a fixed $\boldsymbol{\Phi}$) in the second column. We find this scheme achieves better reconstruction quality because the reconstruction network does not need to constantly adapt to a changing optical encoding (Appendix A.6). We train these microscopes and reconstruction networks on ROIs which are tall cylindrical cutouts of zebrafish with diameter 193 µm and height 200 µm. Sample volumes are imaged on a camera with 256×256 pixels (Figure 3). FourierNet2D has 2 convolution layers (with 99.8% of its kernel parameters in the initial Fourier convolution layer), while UNet2D has 32 convolution layers (kernel parameters approximately uniformly distributed per layer). UNet2D was designed to have a global receptive field. FourierNet2D and UNet2D both have $\sim 4 \times 10^7$ parameters; FourierNet3D has $\sim 6 \times 10^7$ parameters vs. $\sim 10^8$ for UNet3D. The Wiener + UNet method [3] has $\sim 8 \times 10^7$ parameters. Architecture details are in Appendix A.6, A.8.

Simulated camera images (Figure 3) show that the UNet microscope does not make sufficient use of camera pixels, producing only a single view of the volume. We speculate this is due to a local information prior in the small kernels of UNets. Adding a Wiener deconvolution as an approximate inverse before a UNet also does not result in a microscope that makes full use of camera pixels, though better than the UNet alone. The FourierNet microscope uses more camera pixels and performs better than the UNet microscope for reconstruction (Figure 3, quantified in Table 2). Timings in Table

Table 3: ROI specific microscope parameter optimization for 3 types of zebrafish volumes (mean PSNR (top), MS-SSIM (bottom) $\pm$ s.e.m., $n = 10$). Green shows regions of interest.

	Microscope parameters optimized for		
	Type A	Type B	Type C
			
Tested on			
Type A	**49.76** $\pm$ **1.35** **0.998** $\pm$ **0.000**	46.03 $\pm$ 1.33 0.996 $\pm$ 0.001	42.67 $\pm$ 1.14 0.992 $\pm$ 0.002
Type B	35.56 $\pm$ 1.41 0.965 $\pm$ 0.004	**37.34** $\pm$ **0.96** **0.972** $\pm$ **0.003**	35.38 $\pm$ 1.16 0.967 $\pm$ 0.003
Type C	30.87 $\pm$ 1.15 0.912 $\pm$ 0.007	31.48 $\pm$ 0.93 0.920 $\pm$ 0.006	**33.79** $\pm$ **0.90** **0.935** $\pm$ **0.006**

2 are for only the forward pass on a single TITAN Xp GPU; one training iteration on 8 GPUs takes $\sim$0.4 seconds for FourierNet3D, $\sim$0.8 seconds for Wiener + UNet, and $\sim$0.8 seconds for UNet3D (Appendix A.6). Both reconstruction networks must reconstruct from images that have a compressed encoding of 3D information, but the FourierNet2D is clearly more effective than the UNet2D at optimizing this encoding.

3.3 FourierNets outperform UNets for 3D snapshot microscopy volume reconstruction

We can determine which architecture is better for volume reconstruction by choosing fixed microscope parameters and varying the architecture, except for the Wiener + UNet method [3] which is trained in one stage. In Table 2, we compare results using a FourierNet with 3D convolutions (FourierNet3D) and a vanilla UNet with 3D convolutions (UNet3D). UNet3D was also designed to have a global receptive field. Architecture details are in Appendix A.6, A.8.

Reconstruction results in Table 2 compare normalized MSE L_{HNMSE} between the high pass filtered volume and high pass filtered reconstruction, the multiscale structural similarity $\mathrm{MS-SSIM}$ between the true volume and its reconstruction, and finally the peak signal-to-noise ratio PSNR. We also visualize reconstruction results for a volume in the head of a zebrafish in Figure 3. The UNet3D reconstruction networks (using either microscope) fall significantly behind the FourierNet3D reconstruction network in all metrics, despite their global receptive field. The Wiener + UNet [3] network achieves similar $\mathrm{MS-SSIM}$ and PSNR as the FourierNet but a worse L_{HNMSE} due to the inconsistent detail, though the reconstructions are slightly better than the UNet3D for certain regions of the volume.

3.4 Engineered optical encoders can be optimized for a region of interest

To explore the effect of ROI size on optimized $\boldsymbol{\phi}$ and the resulting reconstruction performance, we optimized $\boldsymbol{\phi}$s for three different regions of interest: 1) Type A, ROIs with a short cylinder cutout of 386 µm diameter and 25 µm height, 2) Type B, ROIs with a tall cylinder cutout of 386 µm diameter and 250 µm height, and 3) Type C, ROIs without any cutout of dimension (250 µm z $\times$ 832 µm y $\times$ 832 µm x) (Table 3). All ROI types were imaged with 512×512 pixels on the simulated camera. We then tested reconstruction performance on all combinations of optimized $\boldsymbol{\phi}$ and ROI type, as shown in Table 3 and visualized in Figure 4 for Type B. We include architecture details in Appendix A.6, A.9.

We see in Table 3 that for all types, highest performance is achieved using the phase mask optimized for that particular type. These results show that there is potentially a large benefit to optimizing type-specific optical encoders, which can be easily implemented in a programmable microscope.

3.5 Engineered optical encoders can be implemented on a programmable microscope

We demonstrate that the optical encoders engineered by simulations can be implemented on real hardware using a prototype programmable microscope. Our engineered phase masks were displayed on a spatial light modulator (SLM) located in the pupil plane. We observe a good qualitative match between the features of the simulated and experimental optical encoders (point spread functions) in

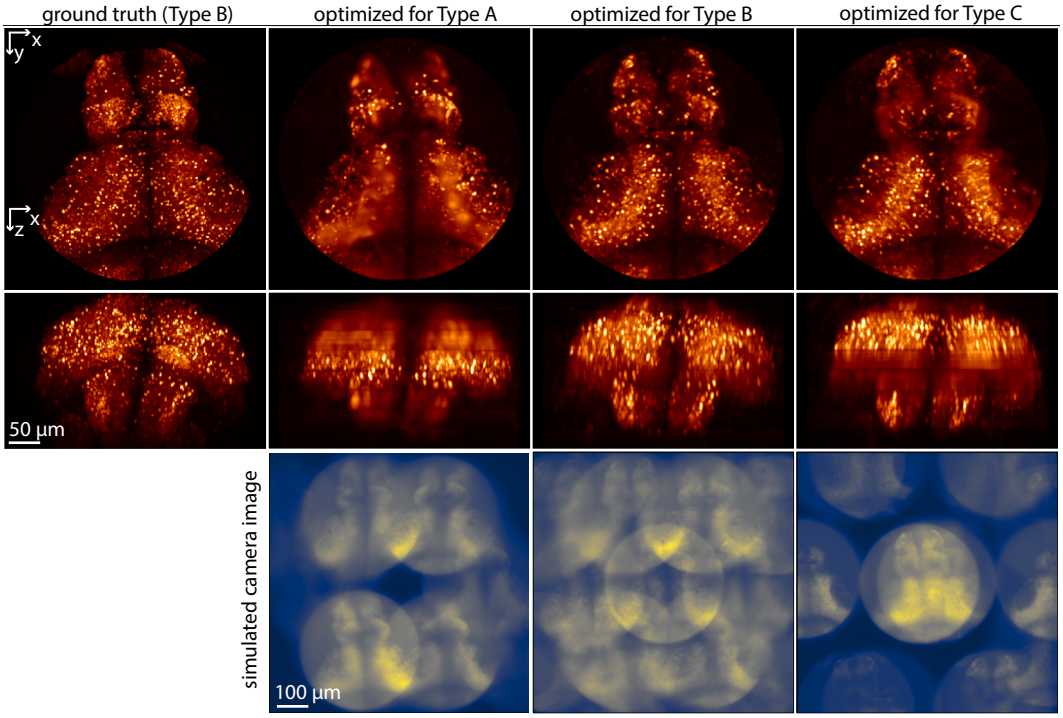


Figure 4: Reconstructed volumes resulting from imaging Type B ROIs by microscopes optimized for Type A, B, C. Imaging Type B ROIs with microscope parameters optimized for Type B ROIs yields the best reconstructions. Top to bottom: xy max projection, xz max projection, simulated 512×512 camera image.

Figure 5, technical details in Appendix A.10. There is also a good qualitative match between the simulated camera images generated by both optical encoders. These results demonstrate the potential for programmable microscopy with spatial light modulators.

We also note that our optimized optical encoders result in pencil-like elements, qualitatively similar to lenslet-based approaches [19, 10, 9]. However as we show in our supplement, our optimized phase masks (Figures 11) appear qualitatively different from lenslet-based phase masks, with different regions of the pupil contributing to the same pencil. This suggests a qualitatively different, and perhaps more light efficient, mechanism for generating high resolution projections of the volume along different axes.

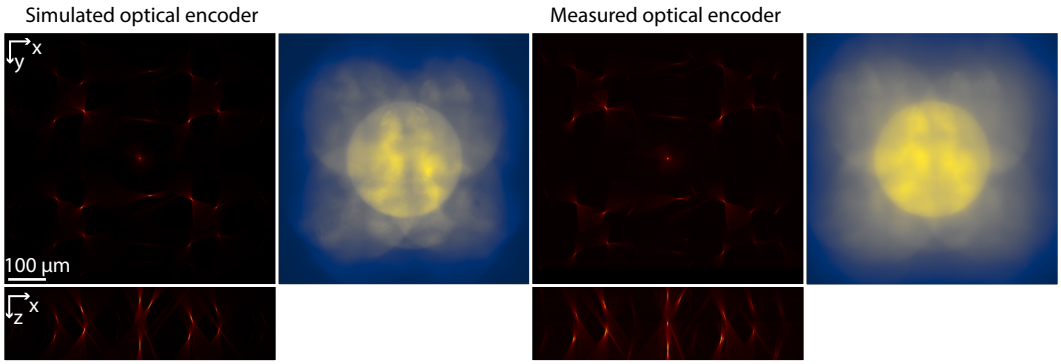


Figure 5: Simulated versus measured optical encoders (point spread functions) with corresponding simulated camera images using an example from our Type B zebrafish dataset. Zoom in for best viewing.

4 Discussion

Summary We have presented the FourierNet architecture as a deep network based decoder for computational imaging with highly non-local optical encoders. We demonstrated the superiority of FourierNets for lensless photography on the DiffuserCam dataset, and also for the end-to-end optimization of highly non-local optical encodings (where UNets fail) for 3D snapshot microscopy. FourierNets are many orders of magnitude faster than traditional iterative reconstruction algorithms [25] and generate higher quality reconstructions by learning image priors. Generally, our global Fourier-domain convolution architecture could be applicable to other problems where global integration of features is necessary, though we have focused on computational imaging applications where physics-based optical encoders induce global mixing of information. Our work solves two important computational problems preventing E2E optimization of large highly non-local optical encoders: the computational complexity of simulating large non-local encoders, and the effective decoding from such encoders. Our contributions are primarily computational, however we also demonstrate the potential for implementing 3D snapshot microscopy with an E2E optimized non-local optical encoder using an SLM-based programmable microscope. And in simulation, we demonstrated the potential for region of interest specific optimization of optical encodings for 3D snapshot microscopy, which can be efficiently implemented in a programmable microscope.

Limitations While we have fully demonstrated that our framework now enables computational E2E optimization of large non-local optical encoders, we do not focus on their hardware implementation. The specific claims regarding the potential benefits of ROI-specific optimization of optical encoders are only evaluated in simulation, and have only been evaluated for whole brain larval zebrafish data.

Reproducibility We train on 8 Quadro RTX 8000 GPUs for the largest experiments, and have described our pre-processing, training, and testing procedures in Appendix A.4, A.6, A.8. We will also release our simulation software, training scripts, and our datasets upon publication.

Broader Impact End-to-end optimization of optics can improve image quality but also enable multiplexed imaging not currently possible with standard optics. Our optical modeling library as well as our neural network architectures will be open-sourced and could enable new experiments in neuroscience through whole-brain imaging at an order of magnitude greater temporal resolution. Our training procedure does require long optimization periods with many GPUs, which poses a carbon footprint and barrier to usage compared to conventional microscopy.

References

[1] J. Chang and G. Wetzstein, "Deep optics for monocular depth estimation and 3d object detection," *arXiv:1904.08601 [cs, eess]*, Apr 2019. arXiv: 1904.08601.

[2] X. Dun, X. Dun, H. Ikoma, G. Wetzstein, Z. Wang, Z. Wang, X. Cheng, X. Cheng, X. Cheng, Y. Peng, and et al., "Learned rotationally symmetric diffractive achromat for full-spectrum computational imaging," *Optica*, vol. 7, p. 913–922, Aug 2020.

[3] H. Ikoma, C. M. Nguyen, C. A. Metzler, Y. Peng, and G. Wetzstein, "Depth from defocus with learned optics for imaging and occlusion-aware depth estimation," in *2021 IEEE International Conference on Computational Photography (ICCP)*, p. 1–12, May 2021.

[4] Y. Wu, V. Boominathan, H. Chen, A. Sankaranarayanan, and A. Veeraraghavan, "Phasecam3d — learning phase masks for passive single view depth estimation," in *2019 IEEE International Conference on Computational Photography (ICCP)*, p. 1–12, May 2019.

[5] Y. Hua, S. Nakamura, M. S. Asif, and A. C. Sankaranarayanan, "Sweepcam — depth-aware lensless imaging using programmable masks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, p. 1606–1617, Jul 2020.

[6] E. Nehme, D. Freedman, R. Gordon, B. Ferdman, L. E. Weiss, O. Alalouf, T. Naor, R. Orange, T. Michaeli, and Y. Shechtman, "Deepstorm3d: dense 3d localization microscopy and psf design by deep learning," *Nature Methods*, vol. 17, p. 734–740, Jul 2020.

[7] H. Ikoma, T. Kudo, Y. Peng, M. Broxton, and G. Wetzstein, "Deep learning multi-shot 3d localization microscopy using hybrid optical-electronic computing," *Optics Letters*, vol. 46, p. 6023–6026, Dec 2021.

[8] M. Levoy, R. Ng, A. Adams, M. Footer, and M. Horowitz, "Light field microscopy," in *ACM SIGGRAPH 2006 Papers*, SIGGRAPH '06, p. 924–934, Association for Computing Machinery, Jul 2006.

[9] L. Cong, Z. Wang, Y. Chai, W. Hang, C. Shang, W. Yang, L. Bai, J. Du, K. Wang, and Q. Wen, "Rapid whole brain imaging of neural activity in freely behaving larval zebrafish (danio rerio)," *eLife*, vol. 6, p. e28158, Sep 2017.

[10] F. Linda Liu, G. Kuo, N. Antipa, K. Yanny, and L. Waller, "Fourier diffuserscope: single-shot 3d fourier light field microscopy with a diffuser," *Optics Express*, vol. 28, p. 28969, Sep 2020.

[11] N. C. Pégard, H.-Y. Liu, N. Antipa, M. Gerlock, H. Adesnik, and L. Waller, "Compressive light-field microscopy for 3d neural activity recording," *Optica*, vol. 3, p. 517–524, May 2016.

[12] R. Prevedel, Y.-G. Yoon, M. Hoffmann, N. Pak, G. Wetzstein, S. Kato, T. Schrödel, R. Raskar, M. Zimmer, E. S. Boyden, and et al., "Simultaneous whole-animal 3d imaging of neuronal activity using light-field microscopy," *Nature Methods*, vol. 11, p. 727–730, Jul 2014.

[13] O. Skocek, T. Nöbauer, L. Weilguny, F. Martínez Traub, C. N. Xia, M. I. Molodtsov, A. Grama, M. Yamagata, D. Aharoni, D. D. Cox, and et al., "High-speed volumetric imaging of neuronal activity in freely moving rodents," *Nature Methods*, vol. 15, p. 429–432, Jun 2018.

[14] K. Yanny, N. Antipa, R. Ng, L. Waller, and L. Waller, "Miniature 3d fluorescence microscope using random microlenses," in *Biophotonics Congress: Optics in the Life Sciences Congress 2019 (BODA,BRAIN,NTM,OMA,OMP) (2019), paper BT3A.4*, p. BT3A.4, Optical Society of America, Apr 2019.

[15] N. Wagner, F. Beuttenmueller, N. Norlin, J. Gierten, J. C. Boffi, J. Wittbrodt, M. Weigert, L. Hufnagel, R. Prevedel, and A. Kreshuk, "Deep learning-enhanced light-field imaging with continuous validation," *Nature Methods*, vol. 18, p. 557–563, May 2021.

[16] N. Antipa, G. Kuo, R. Heckel, B. Mildenhall, E. Bostan, R. Ng, and L. Waller, "Diffusercam: lensless single-exposure 3d imaging," *Optica*, vol. 5, p. 1–9, Jan 2018.

[17] F. L. Liu, V. Madhavan, N. Antipa, G. Kuo, S. Kato, and L. Waller, "Single-shot 3d fluorescence microscopy with fourier diffusercam," in *Biophotonics Congress: Optics in the Life Sciences Congress 2019 (BODA,BRAIN,NTM,OMA,OMP) (2019), paper NS2B.3*, p. NS2B.3, Optical Society of America, Apr 2019.

[18] M. S. Asif, A. Ayremlou, A. Veeraraghavan, R. Baraniuk, and A. Sankaranarayanan, "Flatcam: Replacing lenses with masks and computation," in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, p. 663–666, Dec 2015.

[19] K. Yanny, N. Antipa, W. Liberti, S. Dehaeck, K. Monakhova, F. L. Liu, K. Shen, R. Ng, and L. Waller, "Miniscope3d: optimized single-shot miniature 3d fluorescence microscopy," *Light: Science & Applications*, vol. 9, p. 171, Oct 2020.

[20] M. B. Ahrens, M. B. Orger, D. N. Robson, J. M. Li, and P. J. Keller, "Whole-brain functional imaging at cellular resolution using light-sheet microscopy," *Nature Methods*, vol. 10, p. 413–420, May 2013.

[21] Y. Mu, D. V. Bennett, M. Rubinov, S. Narayan, C.-T. Yang, M. Tanimoto, B. D. Mensh, L. L. Looger, and M. B. Ahrens, "Glia accumulate evidence that actions are futile and suppress unsuccessful behavior," *Cell*, vol. 178, pp. 27–43.e19, Jun 2019.

[22] W. Yang and R. Yuste, "In vivo imaging of neural activity," *Nature Methods*, vol. 14, p. 349–359, Apr 2017.

[23] Y. Zhang, M. Rózsa, Y. Liang, D. Bushey, Z. Wei, J. Zheng, D. Reep, G. J. Broussard, A. Tsang, G. Tsegaye, S. Narayan, C. J. Obara, J.-X. Lim, R. Patel, R. Zhang, M. B. Ahrens, G. C. Turner, S. S.-H. Wang, W. L. Korff, E. R. Schreiter, K. Svoboda, J. P. Hasseman, I. Kolb, and L. L. Looger, "Fast and sensitive gcamp calcium indicators for imaging neural populations," p. 2021.11.08.467793, Nov 2021.

[24] A. S. Abdelfattah, T. Kawashima, A. Singh, O. Novak, H. Liu, Y. Shuai, Y.-C. Huang, L. Campagnola, S. C. Seeman, J. Yu, J. Zheng, J. B. Grimm, R. Patel, J. Friedrich, B. D. Mensh, L. Paninski, J. J. Macklin, G. J. Murphy, K. Podgorski, B.-J. Lin, T.-W. Chen, G. C. Turner, Z. Liu, M. Koyama, K. Svoboda, M. B. Ahrens, L. D. Lavis, and E. R. Schreiter, "Bright and photostable chemigenetic indicators for extended in vivo voltage imaging," *Science*, vol. 365, p. 699–704, Aug 2019.

[25] K. Monakhova, J. Yurtsever, G. Kuo, N. Antipa, K. Yanny, and L. Waller, "Learned reconstructions for practical mask-based lensless imaging," *Optics Express*, vol. 27, p. 28075–28090, Sep 2019.

[26] Z. Wang, L. Zhu, H. Zhang, G. Li, C. Yi, Y. Li, Y. Yang, Y. Ding, M. Zhen, S. Gao, and et al., "Real-time volumetric reconstruction of biological dynamics with light-field microscopy and deep learning," *Nature Methods*, p. 1–6, Feb 2021.

[27] K. Yanny, K. Monakhova, R. W. Shuai, and L. Waller, "Deep learning for fast spatially varying deconvolution," *Optica*, vol. 9, p. 96–99, Jan 2022.

[28] S. Diamond, V. Sitzmann, F. Heide, and G. Wetzstein, "Unrolled optimization with deep priors," *arXiv e-prints*, vol. 1705, p. arXiv:1705.08041, May 2017.

[29] J. Dong, S. Roth, and B. Schiele, "Deep wiener deconvolution: Wiener meets deep learning for image deblurring," *Advances in Neural Information Processing Systems*, vol. 33, p. 1048–1059, 2020.

[30] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters — improve semantic segmentation by global convolutional network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 1743–1751, IEEE, Jul 2017.

[31] N. Mu, X. Xu, and X. Zhang, "Salient object detection in low contrast images via global convolution and boundary refinement," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, p. 743–751, IEEE, Jun 2019.

[32] O. Rippel, J. Snoek, and R. P. Adams, "Spectral representations for convolutional neural networks," in *Advances in Neural Information Processing Systems* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds.), vol. 28, Curran Associates, Inc., 2015.

[33] E. Adelson and J. Wang, "Single lens stereo with a plenoptic camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, p. 99–106, Feb 1992.

[34] L. Grosenick, M. Broxton, C. K. Kim, C. Liston, B. Poole, S. Yang, A. Andalman, E. Scharff, N. Cohen, O. Yizhar, and et al., "Identification of cellular-activity dynamics across large tissue volumes in the mammalian brain," *bioRxiv*, p. 132688, May 2017.

[35] M. Broxton, *Volume Reconstruction and Resolution Limits for Three Dimensional Snapshot Microscopy*. PhD thesis, Stanford University, Aug 2017.

[36] J. Goodman, *Introduction to Fourier Optics*. Macmillan Learning, 4 ed., 2017.

[37] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), Lecture Notes in Computer Science, p. 234–241, Springer International Publishing, 2015.

[38] A. V. Oppenheim and A. S. Willsky, *Signals and Systems: Pearson New International Edition*. Pearson Education Limited, Jul 2013. Google-Books-ID: ut9oAQAACAAJ.

[39] M. Mathieu, M. Henaff, and Y. LeCun, "Fast training of convolutional networks through ffts," *arXiv:1312.5851 [cs]*, Mar 2014. arXiv: 1312.5851.

[40] N. Vasilache, J. Johnson, M. Mathieu, S. Chintala, S. Piantino, and Y. LeCun, "Fast convolutional nets with fbfft: A gpu performance evaluation," *arXiv:1412.7580 [cs]*, Apr 2015. arXiv: 1412.7580.

[41] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv:1502.03167 [cs]*, Mar 2015. arXiv: 1502.03167.

[42] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv:1607.08022 [cs]*, Nov 2017. arXiv: 1607.08022.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See our summary in Section 4.

(b) Did you describe the limitations of your work? [Yes] We also discuss limitations in Section 4.

(c) Did you discuss any potential negative societal impacts of your work? [Yes] We also discuss broader impacts in Section 4.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [N/A]

(b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Appendix A.4, A.6, A.8. We will also release our simulation and optimization library.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix A.4, A.6, A.8.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See standard error of the mean values in Table 2.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See limitations in Section 4 and Appendix A.6.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] We have cited UNets as [37] and DLMD as [25].

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We will be releasing our simulation and reconstruction network code on GitHub.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See footnote 1 about DiffuserCam dataset. Zebrafish data is our own.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]