DEEP LEARNING FOR SPARSE AND LIMTIED-VIEW PHOTOACOUSTIC TOMOGRAPHY IMAGE RECONSTRUCTION

by

Steven Guan
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Bioengineering

Committee:

_____	Dr. Parag Chitnis, Dissertation Director
_____	Dr. Qi Wei, Committee Chair
_____	Dr. Siddhartha Sikdar, Committee Member
_____	Dr. Vadim Sikolov, Committee Member
_____	Dr. Michael Buschmann, Department Chair
_____	Dr. Kenneth S. Ball, Dean, Volgenau School of Engineering

Date:_____ Fall Semester 2021
George Mason University
Fairfax, VA

Deep Learning for Sparse and Limited-View Photoacoustic Tomography Image Reconstruction

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Bioengineering at George Mason University

by

Steven Guan
Master of Science
University of Virginia, 2015
Bachelor of Science
University of Virginia, 2013

Director: Parag Chitnis, Assistant Professor
Department of Bioengineering

Fall Semester 2021
George Mason University
Fairfax, VA

DEDICATION

This is dedicated to my father, mother, and grandparents who have sacrificed their livelihood and comforts to give me an opportunity to grow and learn in the United States.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

Photoacoustic Imaging ..... PAI
Photoacoustic Tomography ..... PAT
Photoacoustic Microscopy ..... PAM
Acoustic Resolution-Photoacoustic Microscopy ..... AR-PAM
Optical Resolution-Photoacoustic Microscopy ..... OR-PAM
Nanosecond ..... ns
Microseconds ..... µs
Convolutional Neural Network ..... CNN
Fully Dense UNet ..... FD-UNet
Peak Signal-to-Noise Ratio ..... PSNR
Structural Similarity Index ..... SSIM
Post-processing Deep Learning ..... Post-DL
Computed Tomography ..... CT
Magnetic Resonance Imaging ..... MRI
Post-processing Deep Learning ..... Post-DL
Modified Direct Deep Learning ..... mDirect-DL
Pixel-wise Deep Learning ..... Pixel-DL
Maximum Intensity Projection ..... MIP
Dense Dilated UNet ..... FD-UNet
Multi Scale Structural Similarity Index ..... MS-SSIM
Partial Differential Equation ..... PDE
Fourier Neural Operator ..... FNO
Mean Squared Error ..... MSE

ABSTRACT

DEEP LEARNING FOR SPARSE AND LIMITED-VIEW PHOTOACOUSTIC TOMOGRAPHY IMAGE RECONSTRUCTION

Steven Guan, PhD

George Mason University, 2021

Dissertation Director: Dr. Parag Chitnis

Photoacoustic tomography (PAT) is a non-ionizing imaging modality capable of acquiring high contrast and resolution images based on optical absorption at depths greater than traditional optical imaging techniques. PAT has matured as a technology to the stage of transitioning from a laboratory to a clinical setting is possible. This presents a wide variety of practical considerations and limitations with instrumentation and data acquisition. Common challenges include having a limited number of available acoustic detectors and a reduced "view" of the imaging target. Forming an image with classical reconstruction methods from insufficient data often result in images with artifacts that degrade image quality. Advanced methods such as iterative reconstruction can be effective in reducing or removing the artifacts. But these methods are also computationally expensive and might not be appropriate in settings requiring near real-time imaging. In this work, we summarize our efforts in utilizing deep learning to address

the deficiencies of sparse spatial sampling and limited-view detection in PAT image reconstruction. We begin with an introduction to fundamental principles of photoacoustic imaging (Chapter 1). This is followed by a brief introduction to deep learning and summarize commonly used deep learning frameworks for PAT image reconstruction (Chapter 2). Next, we describe a novel convolutional neural network architecture termed Fully Dense UNet for sparse PAT image reconstruction (Chapter 3). We then describe pixel-wise deep learning, a data pre-processing step that seeks to provide a more informative input to the neural network (Chapter 4). Next, we describe a modified network architecture termed Dense Dilated UNet that leverages the benefits of dense connectivity and dilated convolutions for 3D PAT image reconstruction (Chapter 5). We then describe Fourier Neural Networks as a fast and general solver for the photoacoustic wave equation (Chapter 6). Finally, we conclude with a discussion of key challenges in using deep learning for PAT image reconstruction and future work (Chapter 7).

CHAPTER ONE: PHOTOACOUSTIC IMAGING

Section One: Introduction

Photoacoustic imaging (PAI) is a non-invasive hybrid imaging modality that uses optical illumination and ultrasound detection to acquire images of chromophores (i.e., optically absorbing molecules) in biological tissue [1]. Given its unique use of light and sound, PAI has several distinct advantages over purely optical and acoustic imaging methods. Optical imaging has a limited imaging depth (<1 mm) due to optical scattering, while PAI can maintain high resolution imaging up to several centimeters because ultrasound scatters several orders of magnitude less than light. PAI can also acquire images without speckling, a signal dependent noise commonly found in ultrasound images [2], [3]. PAI has been rapidly gaining popularity and have shown great potential for many preclinical and clinical imaging applications such as small animal whole-body imaging, breast and prostate cancer imaging, and image guided surgery [4]–[7]. Multispectral photoacoustic imaging can be used for functional imaging such as measuring blood oxygen saturation and metabolism in biological tissues [8]. With the ability to provide both structural and functional information, photoacoustic imaging can reveal novel insights into biological processes and disease pathologies [9].

Section Two: Signal Generation

The photoacoustic effect is a physical phenomenon that describes the conversion from optical to acoustic energy and is the basis for generating a photoacoustic signal. PAI begins by using a nanosecond pulsed laser to illuminate the biological tissue (Fig. 1.) [10]. The light pulse incident on the biological tissue is scattered throughout the medium and will eventually either leave the tissue or is absorbed by optically absorbing molecules called chromophores. The excited chromophores convert the absorbed optical energy into heat through the process of thermoelastic expansion. This occurs on a timescale (~ns) much shorter than the timescale for a local movement of the tissue (~μs). Therefore, the heating is isochoric meaning the rapid local temperature increase is accompanied by a pressure increase, which ultimately results in the generation of acoustic waves.

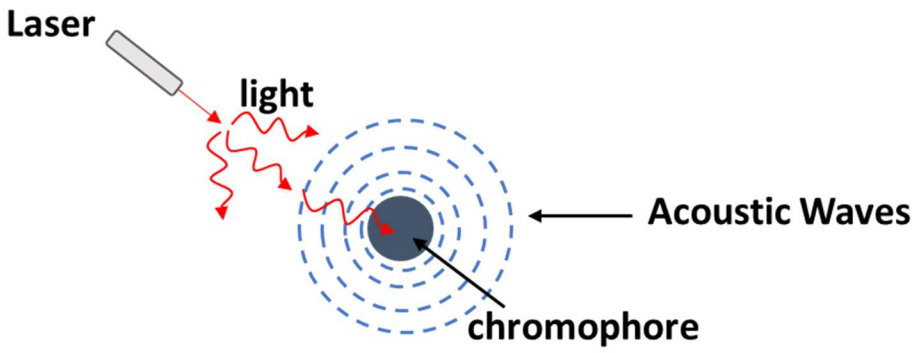

Fig. 1 Process diagram for generating a photoacoustic signal. A chromophore in the tissue absorbs the incident light emitted by the laser and subsequently undergoes thermoelastic expansion. This results in the generation of acoustic waves which then spherically propagates outward with the chromophore acting as a point source.

Photoacoustic imaging readily capitalizes on the rich endogenous and exogeneous optical contrasts available. In biological tissue, chromophores (e.g., hemoglobin, melanin, lipid, and water) exhibit their own characteristic absorption spectra, and their relative quantification can be used to investigate physiological changes (Fig. 2.) [11]. Due to their label-free nature, photoacoustic imaging is well-suited for long term longitudinal monitoring. Various exogeneous contrast agents (e.g., fluorescent dyes and nanoparticles) can also be employed to further enhance imaging specificity, contrast, and depth [12], [13].

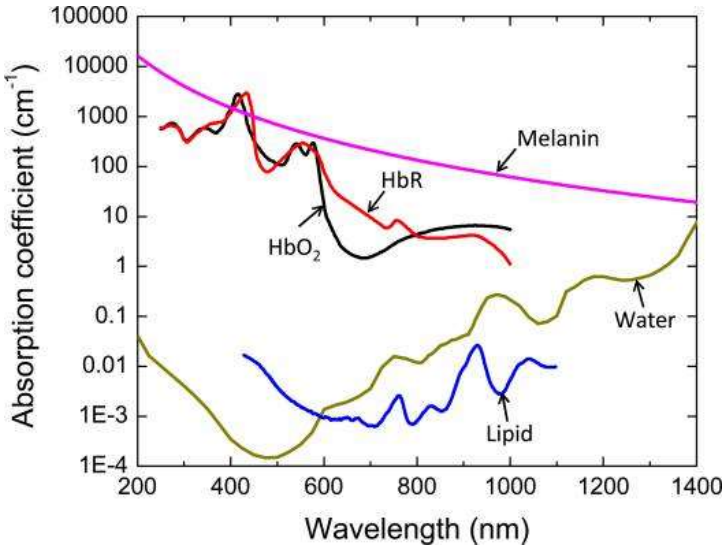

Fig. 2 Absorption coefficient spectra of endogenous tissue chromophores at their typical concentrations in the human body as a function of the incident light's wavelength [11]. The laser wavelength is often tuned to maximize sensitivity for specific chromophores in the tissue.

Assuming negligible thermal diffusion and volume expansion during illumination, the initial acoustic pressure x can be defined as

$$x(r) = \Gamma(r)A(r)$$

where $A(r)$ is the spatial optical absorption function and $\Gamma(r)$ is the Grüneisen coefficient describing the conversion efficiency from heat to pressure [10]. The photoacoustic pressure wave $p(r,t)$ at position r and time t can be modeled as an initial value problem for the wave equation, in which c is the speed of sound [14].

$$(\partial_{tt} - c_0^2\Delta)p(r,t) = 0, \quad p(r,t=0) = x,$$

$$\partial_t p(r,t=0) = 0$$

Acoustic propagation is sensitive to medium properties such as the speed of sound and mass density, and they generally vary throughout the medium. However, the medium is often assumed to be acoustically homogeneous since these variations are often small in soft tissue and is not known in advance. Acoustic absorption is not being considered in this work, but it can be important in some applications [15].

The initial acoustic pressure distribution is related to the optical properties of the tissue and can be written as

$$p_o(r) = \Gamma(r)A(r)$$

where $A(r)$ is the spatial absorption function and $\Gamma(r)$ is the Grüneisen coefficient describing the conversion efficiency from heat energy to pressure [10]. The spatial absorption is a nonlinear function that depends on the optical wavelength, absorption, and scattering through the medium. The initial pressure distribution is often assumed to be initially positive since the absorption of the light and Grüneisen coefficient is positive for most materials.

Section Three: Major Implementations of PAI

PAI has a unique advantage as a multiscale imaging modality that can acquire images at variable spatial resolutions and imaging depths depending on the methods used for optical illumination and acoustic detection [16]. Most major implementations of PAI can be categorized as either photoacoustic microscopy (PAM) or photoacoustic tomography (PAT). PAM typically aims to have an imaging depth of a few millimeters with micron-scale resolution, whereas PAT can have an imaging depth of a few centimeters with either mesoscopic or macroscopic resolution. Choosing an implementation is largely determined by the biomedical application and imaging requirements such as resolution, depth, and speed.

In PAM, the optical illumination and acoustic detection are focused and configured confocally to maximize the PA signal strength [17]. The laser excites tissue predominantly along a line at each scanning position, and the ultrasound transducers receives the PA signals and records the time-of-arrival. A 2D image is formed by raster scanning across one dimension, and depth information is then resolved based on the PA signal's acoustic time-of-flight. PAM can be further categorized as acoustic resolution-PAM (AR-PAM) or optical resolution-PAM (OR-PAM) depending on whether the optical or acoustic focus is finer [18]. AR-PAM utilizes weak optical and tight acoustic focusing with an acoustic lens and is capable of imaging depths up to 3 mm with a resolution of ~20-50 μm [19]. OR-PAM utilizes strong optical and tight acoustic focusing, which enables finer resolutions spanning from hundreds of nanometers to

micrometers [20]. However, the imaging depth in OR-PAM is limited to ~1 mm in biological tissues due to optical scattering.

In PAT, an expanded laser is used to excite the entire tissue sample or a region of interest using wide-field illumination [9]. An array of ultrasound transducers is then used to simultaneously detect the emitted PA waves from multiple view-angles. PAT does not require raster scanning and therefore, is capable of faster cross-sectional and volumetric imaging compared to PAM [21]. Moreover, PAT is capable of greater imaging depths which is essential for many clinical applications such as in human brain and breast imaging. However, PAT generally requires more complex instrumentation and higher computational costs for image reconstruction. Common transducer array geometries such as linear, planar, circular, and semi-circular have been demonstrated for both clinical and animal imaging applications.

Section Four: PAT Signal Measurement

As the generated acoustic waves propagate through the medium, acoustic sensors located along a measurement surface S_o are used to measure a time-series signal [14]. The measurement operator $\mathcal{M}$ acts on $p(r,t)$ restricted to the boundary of the computational domain Ω over a finite time T and provides a linear mapping from the initial pressure x to the measured time-series signal y.

$$y = \mathcal{M}\, p_{|\partial\Omega\times(0,T)} = Ax$$

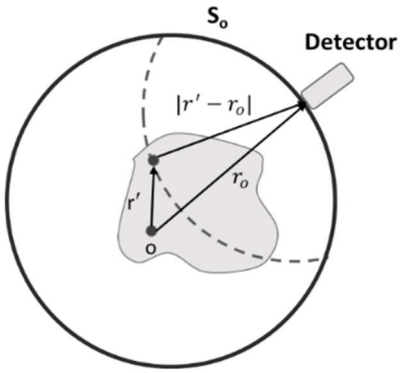

Fig. 3 Diagram illustrating a circular detection geometry of a PAT imaging system. Detector at position r_o on the measurement boundary S_o measures the acoustic pressure emitted from a source located at r'. While illustrated with a circular geometry, the boundary can take any arbitrary shape. Adapted from [14].

The acoustic waves are broadband signals that often have frequencies greater than the range of frequencies detected by the ultrasound transducers used to measure the signal. Furthermore, transducers do have a finite size and are not idealized point detectors which results in the filtering of spatial wavenumbers. To account for these effects, the measurement operator can be written as

$$\mathcal{M} = \mathcal{W}\mathcal{S}$$

where the filtering operator $\mathcal{W}$ accounts for the frequency and wavenumber filtering and the spatial sampling operator $\mathcal{S}$ selects the part of the acoustic field to be measured.

PAT image reconstruction is a well-studied inverse problem that can be solved using analytical solutions, numerical methods (e.g., time reversal), and model-based iterative methods [15], [22]–[25]. With sufficient measured data, the inverse problem is well-posed, and a high-quality PAT image can be reconstructed. To have sufficient data, the imaging system would need a closely spaced array of omnidirectional and broadband point detectors arranged in a geometry such that all rays passing through the imaged

object reach at least one of the detectors. For example, if ideal detectors are positioned on the measurement surface with a spacing of $\lambda_{min}/2$ to satisfy the spatial Nyquist criterion, where λ_{min} is the shortest wavelength generated and the imaging object lies inside the detector array's "visible" region then sufficient data can be measured [26]. Given these strict requirements, data measured in experimental settings often diverge from these ideal conditions leading to an ill-posed inversion problem. PAT images reconstructed from incomplete data often contain artifacts that blur and degrade image quality. A key challenge in PAT reconstruction is to properly account for the deficiencies in the incomplete data and minimize the impact of artifacts.

There are many reasons as to why the incomplete data may be acquired. Below are several reasons commonly found in an experimental setting [27].

1. *Detector responses* are never perfectly broadband or omnidirectional. These characteristics are often selected to achieve sufficient detection sensitivity.
2. *Limited-view detection* meaning the imaging object lies outside of the visible region, and the transducer array has limited "view" or coverage. This can be due to hardware limitations such as using a 2D linear array to image a 3D object or physical limitations restricting the array's coverage.
3. *Undersampling* in space or time to achieve faster data acquisition or due to hardware constraints.

Section Five: Classical Reconstruction Methods

PAT image reconstruction involves an initial acoustic inversion from the measured acoustic time-series data to the initial pressure distribution and an optical inversion to

recover the distribution of optical absorption coefficients [28]. This work is focused on solving the acoustic inversion and are not considering the optical inversion component. In this section, a summary of classical PAT image reconstruction method for the acoustic inversion problem is described.

Back projection was originally used in x-ray tomography image reconstruction, where the measured data is mapped to the image space by projecting the data along a set of lines and summing over all detectors [29]. This concept is also widely used in PAT image reconstruction except the situation in PAT is slightly different. Since the acoustic waves propagate in a spherical manner, the measured data is projected onto spherical shells centered around each detector with a radius based on the signal's time-of-flight and summing over all detector points r_d on the measurement surface S. The back projection operator $\mathcal{A}'$ can be written as,

$$\mathcal{A}'y(r) = \int_S y(r_d, t)_{t=|r-r_d|/c} dS(r_d)$$

where the measured time-series data y is mapped into the image space to reconstruct the image $\hat{x}$. Quality of the reconstructed images can be improved by processing the data before reconstruction and is often referred to as a filtered back projection. The "universal back projection" algorithm is a well-known reconstruction method for PAT, which gives exact reconstructions for common detection geometries such as spherical, cylindrical, and planar [22].

Time reversal is a robust reconstruction method that works well for homogenous and heterogeneous mediums and also for any arbitrary detection geometry [15], [25]. For a measurement surface S surrounding the medium, the acoustic waves generated propagate

outward and are measured as they pass through the surface. After a long period of time T, the acoustic field within the medium will be zero, which is guaranteed by Huygens' principle in 3D homogeneous mediums [30]. A PAT image is formed by running a numerical model of the forward problem backwards in time and transmitting the measured sensor data in a time-reversed order into the medium. Time reversal can be modeled as a time-varying boundary value problems, and the resulting acoustic field at $t = 0$ is the initial acoustic pressure distribution to be recovered.

Iterative methods are commonly employed to remove artifacts and improve image quality. These methods use an explicit model of photoacoustic wave propagation to recover the PAT image x from the measured signal y by solving the following optimization problem using the isotropic total variation (TV) constraint

$$x = \underset{x'}{\operatorname{argmin}} \, || \, y - Ax' ||^2 + \lambda |x'|_{TV}$$

where the parameter $\lambda > 0$ is a regularization parameter [31]–[33]. The TV constraint is a widely employed regularization functional for reducing noise and preserving edges. Framing image reconstruction as a numerical optimization problem provides a flexible framework for tackling the deficiencies of incomplete data. If reconstructing from ideal data, then algorithm will converge to a unique solution. However, if the data is incomplete, then there is not necessarily a unique solution, and the algorithm might be overfitting to noise in the data. The problem of overfitting can be partially by early stopping or including additional terms into the constraint. Iterative reconstruction with a TV constraint works well in the case of simple numerical or experimental phantoms but often leads to sub-optimal reconstructions for images with more complex structures [34]. The main drawback

of iterative methods is that they are computationally expensive due to the repeated evaluations of the forward and adjoint operators

CHAPTER TWO
DEEP LEARNING FOR IMAGE RECONSTRUCTION

Given the wide success of deep learning in classical computer vision tasks such as image segmentation and classification, there is a strong interest in applying similar methods for tomographic image reconstruction problems [35]–[37]. A key factor in the rising popularity of deep learning is that as a data-driven technique, it removes the need for careful feature engineering by an expert and instead directly learns the relevant features needed for the task from a large training dataset. However, these learned features are often highly abstract and incomprehensible to the human eye. Thus, neural networks are frequently treated as a "black box". This is an undesirable trait for biomedical imaging and inverse problems, but recent work has revealed insights into why some network architectures are well-suited for certain tasks and provided justification for the use of deep learning in image reconstruction [38]–[40]. The rising interest in deep learning-based image reconstruction has led to a transition from classical methods to data-driven approaches. Much of this work was in established imaging modalities such as MRI and CT [41]–[44]. In recent years, there has been a growing trend in the literature for PAT image reconstruction for using deep learning to tackle the challenges of incomplete or limited data with the goal of obtaining more accurate and faster image reconstructions than classical methods [45], [46].

Section One: Introduction to Deep Learning

Deep learning or a "deep neural network" is a nonlinear operator that maps an input vector to a target output vector [47], [48]. The network consists of multiple "layers" which is a composition of an affine linear function with learnable parameters and a nonlinearity often referred to as an activation function. A layer $\mathcal{L}$ in the network is defined as

$$\mathcal{L}(h^0) = \varphi(Ch^0 + b) = h^1$$

for a given input vector $h^0 = \{h_j^0\}_{j=1}^J \in \mathbb{R}^J$, where $j \in \mathcal{j} = \{1, \dots, J\}$, an output vector $h^1 = \{h_i^1\}_{i=1}^I \in \mathbb{R}^I$, where $i \in \mathcal{i} = \{1, \dots, I\}$, a linear map given as matrix $C \in \mathbb{R}^{I \times J}$, a vector $b \in \mathbb{R}^I$, and a nonlinear function φ. The term layer typically refers to an operation and its corresponding output except for the "input layer" which refers to the input data without any operation (e.g., the image or measured data).

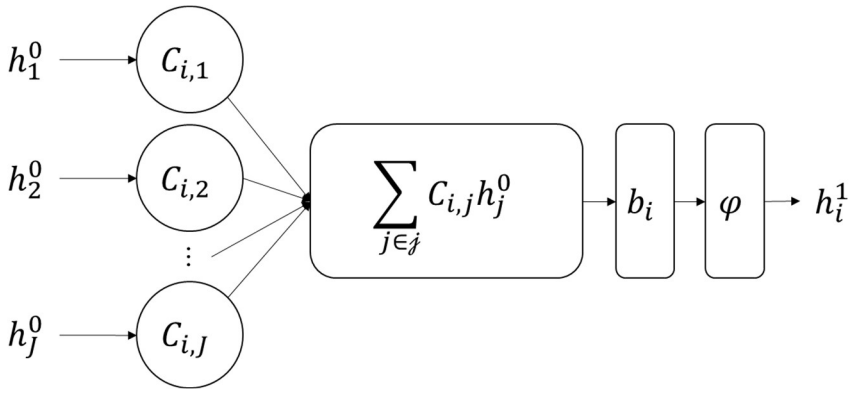

Fig. 4 A single neuron of a layer in an artificial neural network that maps an input vector h^0 to an output vector h^1 using a learned set of coefficients and a non-linear activation function. Adapted from [27].

Within a layer, an individual neuron maps the input vector to one element in the output vector by summing over all input elements of h_0 with a common bias b_i followed by a nonlinearity. Common activation functions used for the nonlinearity are summarized into Table I. The output for the *i*th neuron in a layer is defined as

$$h_i^1 = \varphi\left(\sum_{j \in \mathcal{j}} C_{i,j} h_j^0 + b_i\right) \text{ for each } i \in \mathcal{i}$$

Table 1 Common activation functions used as a nonlinearity in neural networks [49].

Activation Function	$\boldsymbol{\varphi(x)}$	**Values**
Hyperbolic tangent	$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$	(-1, 1)
Sigmoid	$S(x) = \frac{1}{1 + e^{-x}}$	(0, 1)
Rectified linear unit (ReLu)	$R(x) = \begin{cases} 0 \text{ for } x < 0 \\ x \text{ for } x \geq 0 \end{cases}$	$[0, \infty)$
Heaviside function	$H(x) = \begin{cases} 0 \text{ for } x < 0 \\ 1 \text{ for } x \geq 0 \end{cases}$	$[0, 1]$
Signum function	$sgn(x) = \begin{cases} -1 \text{ for } x < 0 \\ 0 \text{ for } x = 0 \\ 1 \text{ for } x > 0 \end{cases}$	[-1, 1]

Fully Connected Layers

The fully connected layer is a commonly used network structure that relates all the input elements to each output element [47], [48]. A network architecture consisting of multiple fully connected layers is the basis for many deep neural networks, especially in those designed for image classification and segmentation. This network can be expressed as the composition of several layers $\mathcal{L}_l$ for $l = 1, \dots, L$

$$h^L = (\mathcal{L}_L \circ \mathcal{L}_{L-1} \circ \dots \circ \mathcal{L}_1)(h^0)$$

The set of trainable parameters θ for such a network includes the matrices and bias vectors, $\theta = \{C^l, C^{l-1}, \dots, C^1, b^l, b^{l-1}, \dots, b^1\}$. For imaging applications, the input image or measured signal needs to be reshaped into a vector before being provided as an input to a fully connected layer.

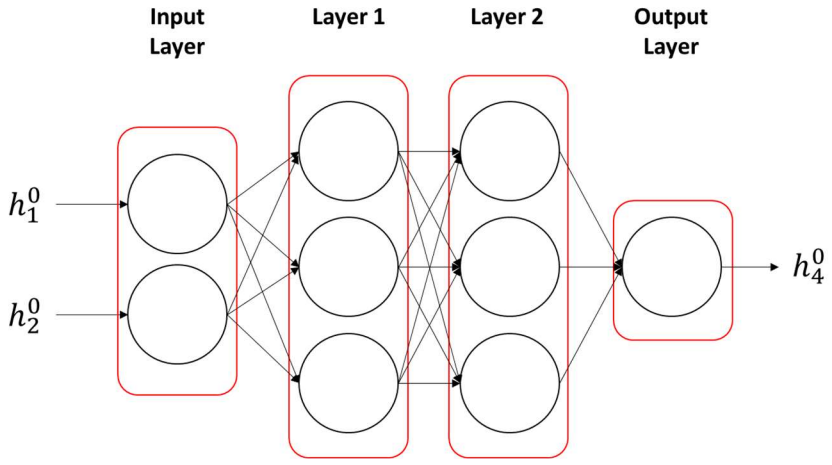

Fig. 5 Diagram for a simple neural network that is comprised of fully connected layers. The network takes an input vector h^0 and computes an output vector h^4 using two hidden layers.

Pixels or voxels in a biomedical image are often related to their nearby neighbors. For example, a neighborhood of pixels with similar values might be part of the same anatomical structure. The spatial relationships between those pixels also represent higher level information such as the size and shape of the structure. These hierarchical features provide critical context for interpreting and analyzing an image. While a fully connected layer can learn to use the spatial information in an image, it is not formulated to do so explicitly nor efficiently.

Convolutional Layers

The convolutional neural network (CNNs) is a type of deep learning architecture that is commonly employed for computer vision tasks. It was inspired by the structure and pattern of neurons in the visual cortex of human and animal brains [50]. Neurons in the visual cortex form a complex sequence to find and focus on small regions in the image like applying local filters over the input. CNNs in a similar manner were designed to efficiently learn local spatial patterns using the convolution operation with a small filter size to identify local features and better exploit the spatial information in an image [51], [52]. The convolution operation is an equivariant operator that can identify similar local features that have been translated or shifted in the image. Convolutional layers are typically more memory and computationally efficient compared to fully connected layers because there are fewer connections between the neurons and only the filter coefficients need to be learned. Thus, CNNs require a smaller number of learnable parameters which greatly simplifies the training process and speeds up the network.

Multiple filters are often used in a single convolutional layer to enable the network to learn many local features. Each filter in the convolutional layer is commonly referred to as a channel. A convolutional layer is defined for each channel as

$$h_i^1 = \varphi\left(\sum_{j\in\mathcal{j}} w_{i,j} * h_j^0 + b_i\right) \text{ for each } i \in \mathcal{i}$$

where $*$ denotes the discrete convolution. The input $h^0 = \{h_j^0 \in \mathbb{R}^{m\times m}\}_{j=1}^{J} \in \mathbb{R}^{m\times m\times J}$ and the output $h^0 = \{h_i^0 \in \mathbb{R}^{m\times m}\}_{i=1}^{I} \in \mathbb{R}^{m\times m\times I}$ are either single or multichannel images, where $j \in \mathcal{j} = \{1, \dots, J\}$ represent the input channels and $i \in \mathcal{i} = \{1, \dots, I\}$ represent the output channels. The set of trainable parameters for a CNN includes the coefficients of the filters w_i and biases b_i.

For many imaging tasks, it is often beneficial to down sample the multichannel inputs as the CNN continues to learn different features. For example, this could be useful in addressing memory constraints by operating on an input with smaller dimensions or increasing the convolutional layer's effective receptive field without requiring additional learnable parameters. The inputs are often down sampled using the mean, median, or maximum filters with a 2x2 kernel, and are termed pooling layers in the CNN. Pooling layers are effective if most of the dominant features or information is retained after the down sampling operation. However, potentially useful information can be lost and reduce the CNN's overall performance.

Section Two: Learning Task Formulation

After defining the neural network architecture, the network is trained to perform a specific task such as image classification, segmentation, or reconstruction. Training the network is an optimization problem with the goal of finding a set of parameters such that the network accurately maps the input to the desired output. In this work, the learning task is defined as an image reconstruction problem, where the network learns to find a mapping from the measured time-series signal to the initial pressure distribution.

The training process is typically formulated as a supervised learning task, where knowledge of the desired output (e.g., ground truth) is known. The neural network learns from paired training examples and updates its parameters via backpropagation based on the error between the network output and corresponding ground truth. Absolute error and mean squared error are commonly used loss functions for training the network. Optimal parameters are learned by minimizing the loss function using optimization strategies such as stochastic gradient descent, RMSprop, and Adam [53], [54].

Supervised learning is the most widely used approach for training a network in PAT image reconstruction literature. However, there are alternative approaches such as semi-supervised and unsupervised learning methods that can train a network using weakly-labeled or unlabeled data [55]. These approaches incorporate an auxiliary measure on the goodness of the reconstructed images to guide the optimization process [56]. They might be useful in scenarios where the training data has only a small number of paired training examples of the input and ground truth. The work presented in this dissertation does not use these alternative approaches for training. Nevertheless, they are

interesting avenues to explore in the future that might provide potential benefits for addressing limitations faced in supervised training such as the network overfitting to the data.

Section Three: Deep Learning Frameworks for PAT Image Reconstruction

Given the numerous applications and measurement configurations for PAT, there has been a wide range of strategies developed for applying deep learning to PAT image reconstruction [27], [57], [58]. These methods are often inspired by key insights in classical methods and can be tailored for a specific imaging task (e.g., denoising). While not intended to be exhaustive, the major frameworks for deep learning in PAT image reconstruction can be roughly grouped depending on how operators for acoustic wave propagation are incorporated into the reconstruction process (Fig. 6).

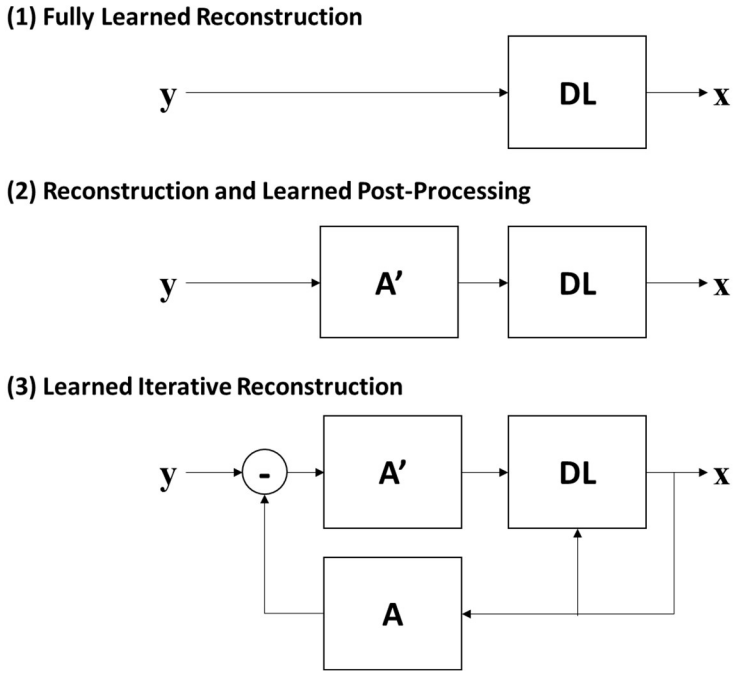

Fig. 6 Summary of frameworks for using deep learning (DL) in the PAT image reconstruction pipeline. The main difference between the frameworks is how the physical model of acoustic wave propagation is used if at all and the input(s) given to the neural network.

Fully Learned Reconstruction

In the fully learned approach or "Direct-DL", the forward and backward operators are not explicitly used in the reconstruction process [59], [60]. The image is formed by directly using a deep neural network to learn the physics required to map the measured time-series data to the image space. This is the most straightforward approach, in which the potentially expensive photoacoustic operators are approximated with a neural network. Depending on the network architecture, the trained network can quickly reconstruct an image with low latency since no explicit model evaluation is required.

In general, this approach relies on fully connected layers to address the issue of nonlocality in the data-to-image transform, where all input elements are related to each

output element. One major limitation is that the fully connected network needs to learn a dense matrix of size $M \times T$ parameters, where M is the total number of pixels and T is the product of the number of detectors and the number of sampling points in time. Due to memory limitations, this approach is typically restricted in application for modestly sized two-dimensions problems. Moreover, the trained network requires consistent dimensions in the data space and image space. Changes in the imaging system configuration such as the number of detectors or the number of time-sampling points would require a new instance of the network to be trained.

Reconstruction and Learned Post-processing

Since the operator for photoacoustic wave propagation is well understood, it would be beneficial to leverage this knowledge rather than having the network start by learning from scratch. In the reconstruction and learned post-processing approach or "Post-DL", this is achieved by using a classical method to provide an initial image reconstruction from the measured time-series data followed by a post-processing step with a neural network to improve image quality [61]. The initial reconstruction is often completed using a fast reconstruction method such as back projection if the goal is to achieve a low latency reconstruction method [22].

Using an initial reconstruction overcomes the inflexibility regarding the acquisition geometry in the fully learned reconstruction since the neural network no longer needs to learn the data-to-image mapping but only an image-to-image mapping. This also allows the post-processing framework to be applied to higher resolution images

and 3D applications since the memory burden of the fully connected layer is removed. Convolutional neural networks are commonly used to learn a restoration operator for an image-to-image mapping. Highly expressive networks with many layers and learnable parameters such as the UNet are typically used to learn the restoration operator [62]. In these networks, the image size is reduced via pooling layers to extract larger spatial features. The extracted coarse features are then subsequently upsampled and combined with previously learned finer features to construct the final image. The main drawback of using highly expressive networks is their tendency to overfit to the training data and failure to generalize to image and artifact types not observed in the training data [63].

Learned Iterative Reconstruction

In the learned post-processing approach, the forward operator is only used once to provide an initial reconstruction. But the forward and its related operators can be used multiple times throughout the reconstruction process to improve the data consistency in the reconstructed images [41], [45]. These approaches are termed learned iterative reconstruction or "model-based" since the neural networks are interlaced with evaluations of the forward, adjoint, and other hand-crafted operators. Repeated use of the forward and adjoint operators enables for more informative inputs to be given to the network to reconstruct a higher quality image. Learned iterative schemes typically outperform other reconstruction approaches but at the cost of a higher computational complexity due to repeated evaluations of the forward and adjoint operators. For each iteration, a neural network with its own set of learnable parameters is trained to update the image with the

goal of minimizing the data consistency term $\mathcal{D}(x;y) = \frac{1}{2}\|y - Ax\|_2^2$. Using a gradient descent scheme for updating the image with a CNN for N steps, this process can be formulated as

$$x^{n+1} = \Lambda_{\theta_n}\left(x^{(n)}, A'\left(Ax^{(n)} - y\right)\right), \quad n = 0, \ldots, N-1$$

where A is the forward operator, A' is the adjoint or a similar hand-crafted operator, and Λ_{θ_n} is a neural network termed the learned updating operator for the nth step. There is an initialization step to map the measured time-series data to the image space. The CNN at the nth step receive the current image to be updated and a gradient image measuring data consistency as inputs.

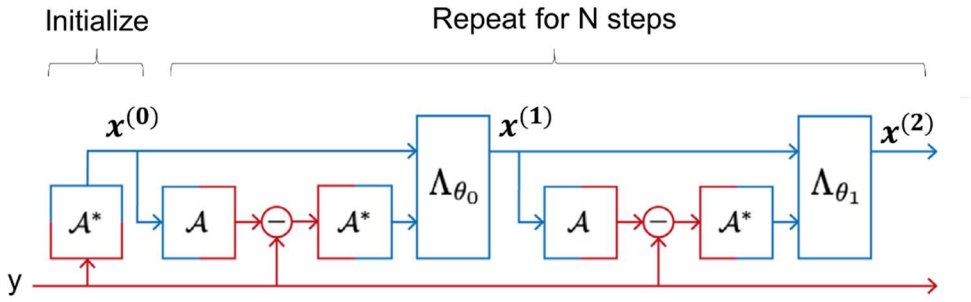

Fig. 7 Process diagram to illustrate the learned iterative approach for the first two iterations and onwards. The measured time-series data y is initially reconstructed into the image x^0. A convolutional neural network Λ_θ iteratively updates the image using the current image and a gradient image as inputs. The colors red and blue refer to the data and image spaces, respectively. Adapted from [27].

CHAPTER THREE
FULLY DENSE UNET FOR 2D PAT

To the best of our knowledge, the first work describing the application of deep learning for photoacoustic tomography image reconstruction was published by Antholzer et al. in 2017 on the pre-print server ArXiv [64]. This preliminary work provided an initial demonstration of the learned post-processing approach and evidence that deep learning is a powerful and effective tool for artifact removal. It also served as a key inspiration for exploring different deep learning methodologies and frameworks. Key challenges identified in their work were the issues of overfitting and data mismatch between the training and testing data. Deep learning networks often perform well on images like the training data and fail to generalize to other images not in the training data. These problems are ubiquitous to all applications of deep learning and machine learning. In this work, the Fully Dense UNet addressed these challenges by incorporating dense connectivity into the well-known UNet architecture [65], [66]. *In Silico* experiments comparing multiple reconstruction methods were performed on a variety of simulated phantoms. Results were published in the Journal of Biomedical Health and Informatics [63].

Section One: Introduction and Motivation

A common challenge faced in PAT is that the acoustic waves can only be sparsely sampled in the spatial dimension. Each discrete spatial measurement requires its own detector, and it may be infeasible to build an imaging system with a sufficiently large

number of detectors due to practical and physical limitations [32], [67], [68]. Reconstructing sparsely sampled data using standard methods result in low quality images with severe artifacts. Iterative reconstruction methods can be used to reduce artifacts and improve image quality by incorporating prior knowledge such as smoothness, sparsity, and total variation constraints into the reconstruction process [31], [32], [69], [70]. However, selecting appropriate constraints can be a challenging task, especially for images with complex spatial structures. Furthermore, iterative methods can be time consuming because they require repeated evaluations of the forward and adjoint operators.

Many applications of deep learning for sparse tomographic image reconstruction follows a post-processing approach, where an initial corrupted image is first reconstructed from the sensor data using a simple inversion step and then a CNN is applied as a post-processing step for removing artifacts and improving image quality. This approach has been successfully applied to CT, MRI, and PAT and shown to achieve comparable image quality to iterative methods [44], [46], [64], [71].

In this work, we follow the post-processing approach and propose a modified CNN architecture termed Fully Dense UNet (FD-UNet) for removing artifacts in 2D PAT images reconstructed from sparse data. The FD-UNet incorporates dense connectivity into the contracting and expanding paths of the UNet CNN architecture. Dense connectivity mitigates learning redundant features and enhances information flow allowing for a more compact and superior CNN [65], [72], [73].

The UNet is the most widely used CNN architecture for applying deep learning with the post-processing approach in sparse tomographic image reconstruction [64], [71], [74]. It has many properties well-suited for artifact removal such as its use of multilevel decomposition and multichannel filtering [44]. Moreover, it has been demonstrated to perform comparatively well to iterative methods for sparse PAT image artifact removal on synthetic and experimental data [64], [74]. We build upon previous work and improve the post-processing approach by incorporating a recent advancement in CNN architecture design, namely dense connectivity, to achieve a CNN with superior performance. Compared to previous UNet implementations, we also apply batch normalization to accelerate the training process [75], [76].

A UNet with dense connectivity termed "DD-Net" has been previously used for sparse-view CT reconstruction and was shown to outperform iterative methods [77]. While the FD-UNet also uses dense connectivity, there are several differences in implementation. 1) The DD-Net includes dense connectivity only in the contracting path of the UNet. Whereas the FD-UNet includes dense connectivity in both the contracting and expanding paths. This strategy enables the benefits of dense connectivity to be leveraged throughout the entire network. 2) In the DD-Net, the dense block "growth rate" hyperparameter remains constant throughout the network. In the FD-UNet, this hyperparameter is updated throughout the CNN to improve computational efficiency. To the best of our knowledge, this is the first work applying the UNet with dense connectivity for removing artifacts in sparse PAT image reconstruction.

Section Two: Methods

As shown in Fig. 8., the sparsely sampled acoustic pressure is initially reconstructed using TR into an image $\boldsymbol{X}$ containing artifacts. The CNN is then applied to correct the undersampling artifacts in image $\boldsymbol{X}$ to obtain an approximately artifact-free image $\boldsymbol{Y}$. This task can be formulated as a supervised learning problem, in which the goal is to learn a restoration function that maps an input image $\boldsymbol{X}$ to the desired output image $\boldsymbol{Y}$ [64]. Other reconstruction methods can be used in place of TR to reconstruct the initial artifact image $\boldsymbol{X}$ from sensor data. TR was chosen for this work because it can be easily adapted for any sensor configuration, provides a good initial reconstruction, and is computationally inexpensive relative to iterative methods.

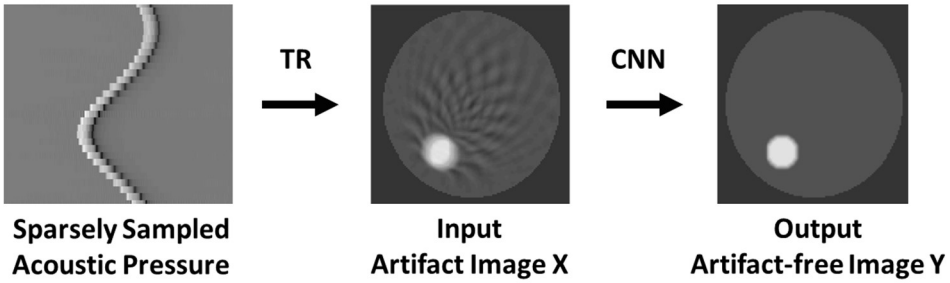

Fig. 8 Deep learning framework for 2D PAT image reconstruction. The sparsely sampled acoustic pressure is reconstructed into an image containing artifacts using time reversal. A CNN is applied to the artifact image *X* to obtain an approximately artifact free image *Y*.

Proposed FD-UNet Architecture

As seen in Fig. 9., the input image $\boldsymbol{X}$ undergoes a multilevel decomposition in the contracting path of the FD-UNet, where the spatial dimensions of the feature maps are repeatedly reduced via a max-pooling operator [44], [62], [78]. This strategy enables the CNN to efficiently learn local and global features relevant for artifact removal at various spatial scales [79]. In the following expanding path, the learned feature-maps are spatially upsampled via a deconvolution operator and combined to produce an output image $\boldsymbol{Y}$ with the same dimensions as the input image $\boldsymbol{X}$. Deconvolution can be thought as the reverse of convolution and is essentially a transposed convolution.

For each spatial level, s, in the FD-UNet, a dense block with a growth rate, k_s, is used to learn several feature-maps, f_s. Initial values for k_1 and f_1 are hyperparameters defined by the user. k_s is updated at each spatial level so that all dense blocks in the FD-UNet have the same number of convolutional layers to maintain computational efficiency. In our implementation, $k_s = 2^{s-1} \times k_1$ and $f_s = 2^{s-1} \times f_1$. Where the FD-UNet use dense blocks, the UNet have instead a sequence of two 3x3 convolution operations to learn feature-maps [64].

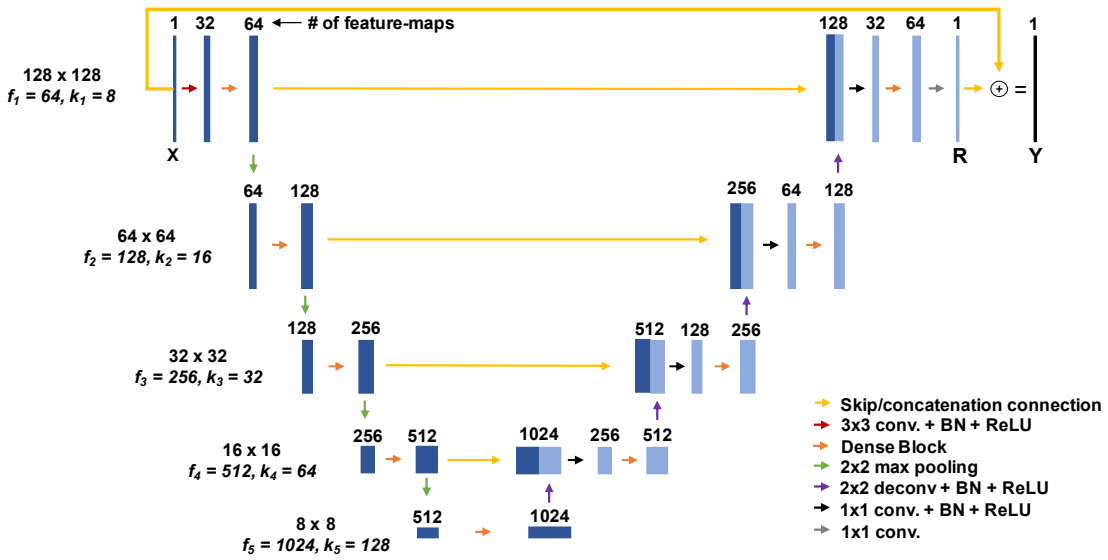

Fig. 9 Proposed FD-UNet architecture that incorporates dense connectivity [26] into the expanding and contracting path of the U-Net [19]. Hyperparameters for the illustrated architecture are $k_1 = 8$ and $f_1 = 64$ for an input image X of size 128x128 pixels.

After each deconvolution operation, the upsampled feature-maps are concatenated channel-wise with feature-maps of similar size from the contracting path. These concatenation connections allow higher resolution features learned earlier in the network to be used in the upsampling process. However, this results in $2f_s$ feature-maps and cannot be reduced to the desired f_s feature-maps using a dense block. To address this issue, the concatenated feature-maps are first reduced to $f_s/2$ feature-maps using a 1x1 convolution prior to each dense block in the expanding path.

In a dense block, earlier convolutional layers are connected to all subsequent layers via channel-wise concatenation [65], [72]. This means that the input to each layer in a dense block is the outputs from all previous layers concatenated together. Essentially, each layer learns additional feature-maps based on the "collective

knowledge" gained by previous layers. This strategy increases the network's representational power through feature reuse. Features learned in earlier layers are passed forward and removes the need to learn redundant features and promotes learning a diverse set of features.

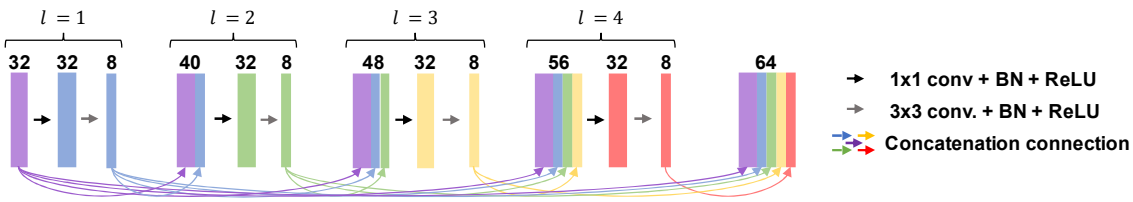

Fig. 10 Four layered dense block with $k_1 = 8$ and $F = 32$. Feature-maps from previous layers are concatenated together as the input to following layers.

Furthermore, dense connectivity allows for deeper networks. For example, the FD-UNet has 82 convolution and deconvolution layers while the UNet has 23 layers. As the depth of the network increases, gradient information passes through many layers and can be lost before it reaches the earlier layers in a network leading to the vanishing gradient problem. Previous networks (e.g. ResNets and Highway Networks) addresses this problem by introducing short paths from earlier to later layers [80], [81]. Dense connectivity follows a similar principle but introduces many more connections to allow for gradient information to be efficiently backpropagated. This mitigates the vanishing gradient problem and allows for the network to be more easily trained.

As seen in Fig. 10., the ℓ^{th} layer in the dense block has an output with k_s feature-maps and an input with $F + k_s \times (\ell - 1)$ feature-maps, where F is the number of

feature-maps of the initial input to the dense block. Features are learned through a sequence of a 1x1 and 3x3 convolution with batch normalization and rectified linear unit (ReLU) activation function [75], [76]. The 1x1 convolution is included to improve computational efficiency by reducing the input size to F feature-maps prior to the more computationally expensive 3x3 convolution. Then k_s features maps are learned from the reduced input using a 3x3 convolution. The final output of the dense block is the concatenation between the input and outputs from all dense block layers.

The proposed CNN architecture utilizes residual learning by adding a skip connection between the input and output [80], [82]. In residual learning, the CNN learns to map the input image $\boldsymbol{X}$ to a residual image $\boldsymbol{R} = \boldsymbol{Y} - \boldsymbol{X}$ and then recovers the target artifact-free image $\boldsymbol{Y}$ by adding the residual $\boldsymbol{R}$ to the input $\boldsymbol{X}$. Residual learning is shown to mitigate the vanishing gradient problem. The residual $\boldsymbol{R}$ often has a simpler structure than the original image and is easier for the CNN to learn [71].

Synthetic Data for Training and Testing

Synthetic training and testing data is created using k-Wave, a MATLAB toolbox for simulating photoacoustic wave fields [83]. For each dataset generated, an initial photoacoustic source with a grid size of 128x128 pixels is defined. The medium is assumed to be non-absorbing and homogenous with a speed of sound of 1500 m/s. The sensor array has N detectors equally spaced on a circle with a radius of 60 pixels. Built-in functions of k-Wave are used to simulate sparse sampling of photoacoustic pressures.

The TR method is then used to reconstruct an initial image containing artifacts from the sparsely sampled data.

Datasets are generated from three different synthetic phantoms (circles, Shepp-Logan, and vasculature) and an anatomically realistic vasculature phantom created from experimentally acquired micro-CT images of mouse brain vasculature. The phantoms are used to define an initial photoacoustic pressure source in k-Wave for creating simulated PAT images.

The circles dataset is comprised of simple phantoms that contain up to five circles with equal magnitude. The center location and radius for each circle are chosen randomly from a uniform distribution. This protocol is used to initially create a total of 1200 circles phantom images. We employed four-fold cross validation by dividing the images into four sets of a 1000 training images and 200 testing images. The images are used to initialize the photoacoustic pressure distribution to created simulated PAT image datasets for three levels of sampling sparsity (10, 15, and 30 detectors).

The Shepp-Logan and synthetic vasculature datasets are created using a data augmentation strategy. Training and testing images are procedurally generated from an original image with a size of 340x340 pixels for each phantom. Downsampled versions of these initial phantom images are shown as ground truth in Fig. 8. New images are created using the following steps. First, scaling and rotation is applied to the original image with a randomly chosen scaling factor (0.5 to 2) and rotation angle (0-359 degrees). Then a 128x128 pixels sub-image is randomly sampled from the transformed image. Finally, the sub-image is translated with a randomly selected vertical and horizontal shift (0-10

pixels) via zero-padding. Data augmentation allows for large sets of images with similar but different features to be easily created [84]. This strategy is used to generate a testing and fine-tuning dataset with 200 and 100 images, respectively, for each synthetic phantom. PAT images are then simulated using k-Wave with a sensor array of 30 detectors.

The anatomically realistic vasculature phantom is derived from a 3D volume of mouse brain vasculature that was experimentally acquired using micro-CT [85]. The original volume had a size of 260x336x438 pixels. The Frangi vesselness filter is applied to suppress background noise and enhance vessel-like features in the volume [86]. New images are created from the filtered volume following a similar data augmentation procedure as described for the synthetic phantoms. However, a 128x128x128 pixels sub-volume is instead randomly sampled from the transformed volume and is used to create a maximum intensity projection image by applying the max operator along the third dimension. Only a testing dataset with 200 images is generated from the mouse brain vasculature phantom. The corresponding training dataset with 1000 images is instead generated from the synthetic vasculature phantom. To create more complex synthetic images for training, the outputs from multiple iterations (up to five) of the data augmentation process are summed together. This enables the synthetic training images to have more a complex network structure with varying vessel sizes and orientation. PAT images of the synthetic and realistic vasculature phantoms are simulated at various levels of sampling sparsity (15, 30, and 45 detectors).

Deep Learning Implementation

The CNNs are implemented in Python 3.6 with TensorFlow v1.7, an open source library for deep learning [87]. Training and evaluation of the network is performed on a GTX 1080Ti NVIDIA GPU. The CNNs are trained for 10,000 iterations using a mean squared error loss function, learning rate of 1e-4, and a mini-batch size of three images.

Section Three: Results

The UNet and FD-UNet are compared over several experiments to determine if dense connectivity enables for more artifacts to be removed and hence an image with higher quality to be recovered. Image reconstruction quality is quantified using the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [88]. PSNR provides a global measurement of image quality whereas SSIM measures the similarity between local patterns of pixel intensities.

Circles Dataset

In this initial experiment, the CNNs are both trained and tested using the circles dataset. This represents an ideal data scenario where the training and testing data are well-matched meaning the CNN had an opportunity to learn almost all the features needed from the training data to perform well on the testing data. This ideal scenario provides a starting point for comparing the performance of the CNNs without limitations from data-related issues. Since the training and testing are derived from the same phantom in this experiment, four-fold cross validation is employed to increase confidence in the results observed.

Table 2 Average PSNR and SSIM for 2D circles dataset (30 Sensors)

	$f_1 = 8$ $k_1 = 1$	$f_1 = 16$ $k_1 = 2$	$f_1 = 32$ $k_1 = 4$	$f_1 = 64$ $k_1 = 8$
TR	32.48 ± 3.52 *0.75 ± 0.07*			
UNet	33.77 ± 4.18 *0.78 ± 0.12* 487K (0.94)	34.48 ± 4.19 *0.79 ± 0.12* 1.9M (1.05)	34.70 ± 4.54 *0.79 ± 0.12* 7.8M (1.55)	34.84 ± 4.48 *0.79 ± 0.12* 31M (2.94)
FD-UNet	39.35 ± 3.19 *0.84 ± 0.08* 151K (0.80)	41.45 ± 3.28 *0.85 ± 0.07* 600K (0.91)	43.05 ± 3.27 *0.86 ± 0.07* 2.4M (1.4)	44.84 ± 3.42 *0.87 ± 0.07* 9.4M (2.78)

f_1 and k_1 are CNN hyperparameters. k_1 is only applicable to the FD-UNet. For each row, the following metrics are reported: PSNR, SSIM in italics, number of trainable parameters, and evaluation time in milliseconds for a single image in parenthesis.

The CNNs' potential in learning to remove artifacts are evaluated by varying the hyperparameters f_1 (initial feature-maps learned) and k_1 (initial dense block growth rate). Increasing f_1 results in a wider CNN with more representational power and typically better performance. Results for the FD-UNet and UNet with varying model complexities for the circles dataset are shown in Table 2 and Fig. 11a. As expected, the initial TR reconstruction has severe artifacts and the lowest average PSNR and SSIM. Applying either CNN generally results in an improved and near artifact-free image. However, the FD-UNet outperforms and is more consistent in removing artifacts than the UNet. As seen in Fig. 11b., the FD-UNet removes majority of the artifacts but the UNet fails to remove artifacts on the boundary of the top-left circle and in the background. For all images in the testing dataset, there are no instances of the UNet outperforming the FD-UNet.

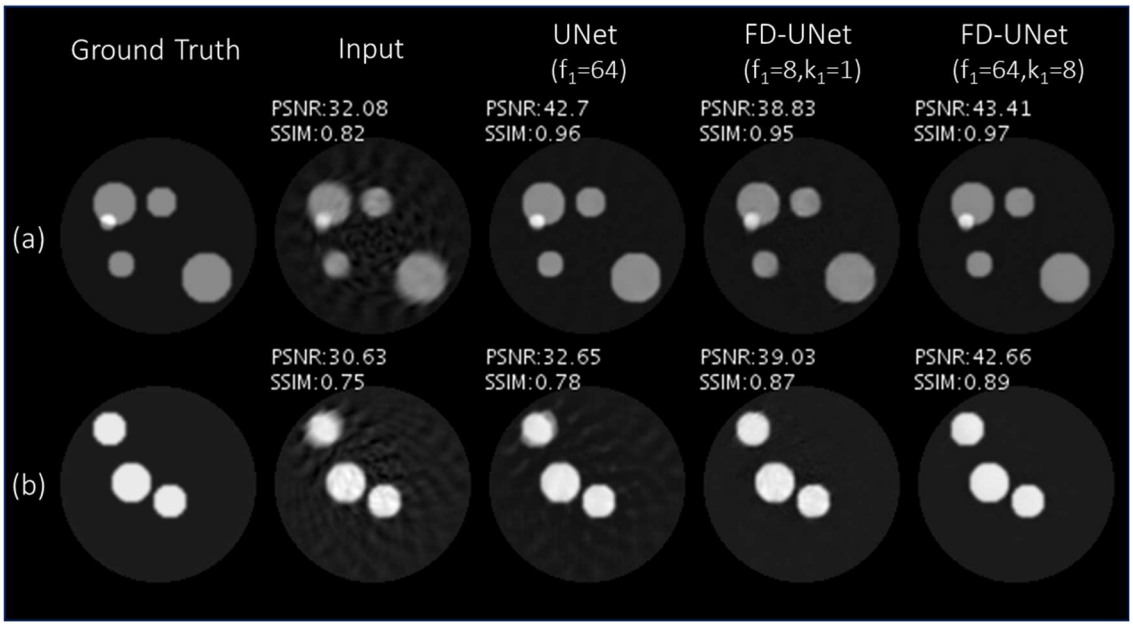

Fig. 11 Reconstructed circles images using TR, UNet, and FD-UNet with varying hyperparameters. (a) both CNNs recover a near artifact-free image. (b) example of the UNet reconstruction with residual background artifacts and the top-left circle has a distorted boundary.

Dense connectivity improves model parameter efficiency and allows for a more compact CCN with better performance. As seen in Table 2, the FD-UNet requires fewer parameters (about a third) and has a higher average PSNR and SSIM compared to the UNet for each set of hyperparameters tested. The CNNs have similar average evaluation times with the FD-UNet being only slightly faster by a fraction of a millisecond. In the FD-UNet, a dense block is used in place of the two 3x3 convolutions in the UNet. While the dense block has eight different convolutional layers (four 1x1 and four 3x3), the input and output of each convolutional layer are relatively smaller. Thus, the convolutional layers in the dense block are computationally cheaper than those in the UNet resulting in the two CNNs having similar evaluation times.

Interestingly, the most compact FD-UNet ($f_1 = 8, k_1 = 1$) with fewer parameters and features learned outperforms the more complex UNet ($f_1 = 64$). This demonstrates that the FD-UNet, despite learning fewer features, is learning more relevant ones for artifact removal. In general, both CNNs have improved performance as f_1 and model complexity is increased. However, these improvements are diminishing because larger CNNs are more difficult to train and prone to overfitting. As seen in Fig. 12., the CNNs are trained for a total of 10,000 iterations but converge to a maximum by 8,000 iterations. The UNet loss appears to be more volatile compared to the FD-UNet loss.

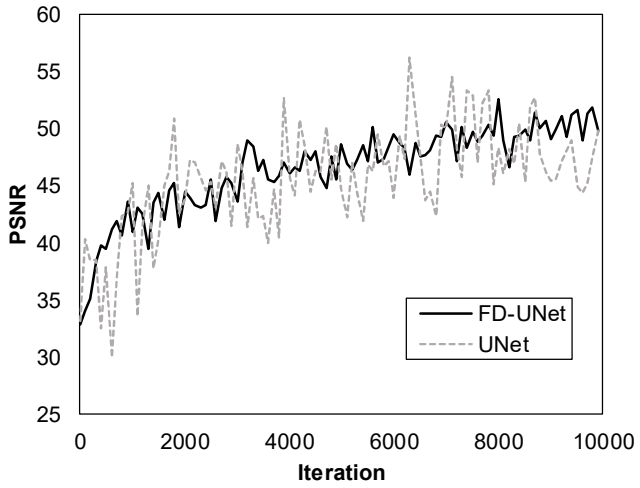

Fig. 12 Training loss in PSNR during the training phase for the FD-UNet ($f_1 = 64, k_1 = 8$) and UNet ($f_1 = 64$) on the circles training dataset (N=30 sensors).

The CNNs' ability to remove artifacts under varying levels of sampling sparsity are also evaluated. The goal of this experiment is to determine the extent of artifact severity that can be removed by each CNN. For each level of sampling sparsity, the CNNs are trained and tested on the corresponding datasets.

Table 3 Average PSNR and SSIM under varying sampling sparsity levels

# of Detectors	10	15	30
TR	24.86 ± 3.18 *0.70 ± 0.05*	27.30 ± 3.15 *0.72 ± 0.06*	32.48 ± 3.52 *0.75 ± 0.07*
UNet	24.69 ± 3.79 *0.72 ± 0.11*	27.26 ± 3.94 *0.76 ± 0.11*	34.84 ± 4.48 *0.79 ± 0.12*
FD-UNet	32.59 ± 4.36 *0.83 ± 0.07*	38.10 ± 4.20 *0.86 ± 0.07*	44.84 ± 3.42 *0.87 ± 0.07*

For each row, PSNR is shown as normal text on top while SSIM is shown as italicized text on the bottom. The CNN hyperparameters used are FD-UNet ($f_1 = 64, k_1 = 8$) and UNet ($f_1 = 64$)

Results for the FD-UNet and UNet for different levels of sampling sparsity are shown in Table 3. As expected, decreasing the number of detectors used to sample the acoustic pressure results in more severe artifacts and a lower average PSNR and SSIM. The FD-UNet has a higher average PSNR and SSIM compared to the UNet for all levels of sampling sparsity tested. Reconstructed phantom images under different levels of sampling sparsity are shown in Fig. 13. When using 30 detectors, both CNNs perform well in removing artifacts from images reconstructed. At a sparser sampling level using 15 detectors, the FD-UNet recovers higher quality images than the UNet. For example, the boundaries of the circles as indicated by the red arrows in Fig. 13b. are blurred together in the UNet reconstruction but can be clearly distinguished in the FD-UNet reconstruction. Both CNNs are unable to reliably reconstruct the circles' boundaries at sparsity level using 10 detectors. Interestingly, the FD-UNet is able recover a reconstruction with a higher SSIM from a more corrupted initial image (10 detectors) than the UNet can from an initial image with less artifacts (30 detectors).

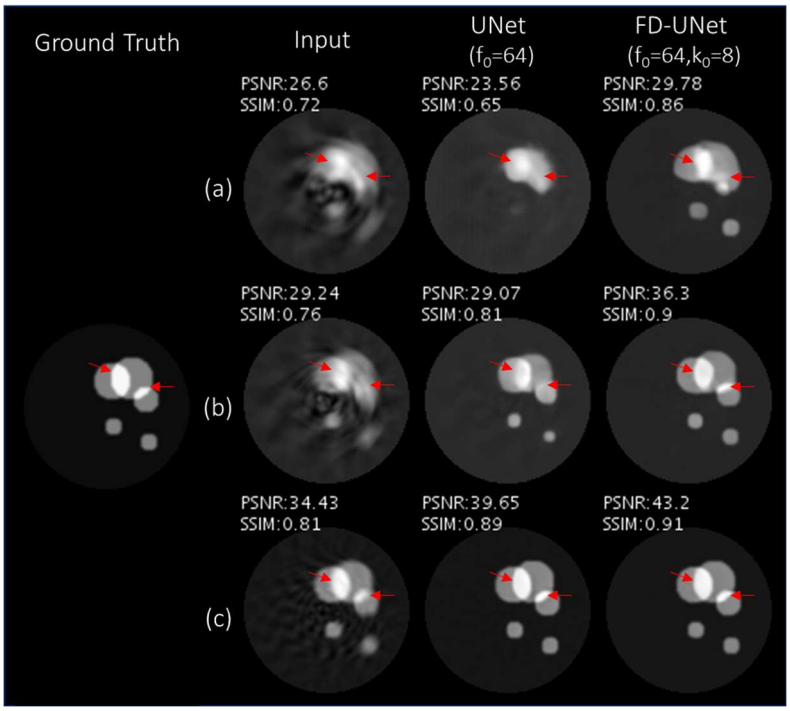

Fig. 13 Reconstructed circles images under different levels of sampling sparsity using (a) 10, (b) 15, and (c) 30 detectors. The red arrows point to a boundary that is blurred at more sparse sampling levels.

Shepp-Logan and Vasculature Phantom Dataset

In the second experiment, the CNNs are initially trained on the circles dataset and tested on the Shepp-Logan and synthetic vasculature data. This represents a scenario in which the training and testing data are not necessarily well-matched. The circles and Shepp-Logan phantoms have many similar circular-like features and are well-matched. However, the circles and synthetic vasculature phantom have significantly different features and are not well-matched. After initially training on the circles dataset, the CNNs are further trained for 5,000 iterations on either the Shepp-Logan or synthetic vasculature fine-tuning dataset. The purpose of this experiment is to evaluate the CNN's performance and ability to generalize when the training and testing datasets are not well-matched.

Furthermore, the feasibility of training on a large poorly matched dataset and a smaller well-matched dataset is explored.

Table 4 Average PSNR and SSIM for Shepp-logan and Vasculature phantom dataset (30 Detectors)

	Shepp-Logan		Vasculature	
	Initial	Fine-tuned	Initial	Fine-tuned
TR	32.50 ± 1.53 *0.87 ± 0.03*		24.79 ± 2.86 *0.66 ± 0.06*	
UNet	31.69 ± 1.19 *0.93 ± 0.03*	36.23 ± 2.46 *0.95 ± 0.04*	24.40 ± 2.93 *0.66 ± 0.06*	25.96 ± 2.85 *0.70 ± 0.11*
FD-UNet	30.81 ± 0.97 *0.94 ± 0.01*	38.24 ± 1.69 *0.97 ± 0.01*	25.27 ± 2.16 *0.70 ± 0.05*	31.30 ± 2.24 *0.82 ± 0.07*

For each row, PSNR is shown as normal text on top while SSIM is shown as italicized text on the bottom. The CNN hyperparameters used are FD-UNet ($f_1 = 64, k_1 = 8$) and UNet ($f_1 = 64$)

Results for the FD-UNet and UNet with and without fine-tuning for the Shepp-Logan and synthetic vasculature datasets are shown in Table 4 and Fig. 14. Both CNNs without fine-tuning have comparable performance and recover a high-quality albeit blurred reconstruction of the Shepp-Logan phantom as seen in Fig. 14a. However, they are not able to perform as well in the case of the of the synthetic vasculature phantom as seen in Fig. 14b. The general structure of the vessels can be clearly seen but appear to have circular-like features like the circles phantom training dataset. The FD-UNet does perform slightly better and removes more of the background artifacts.

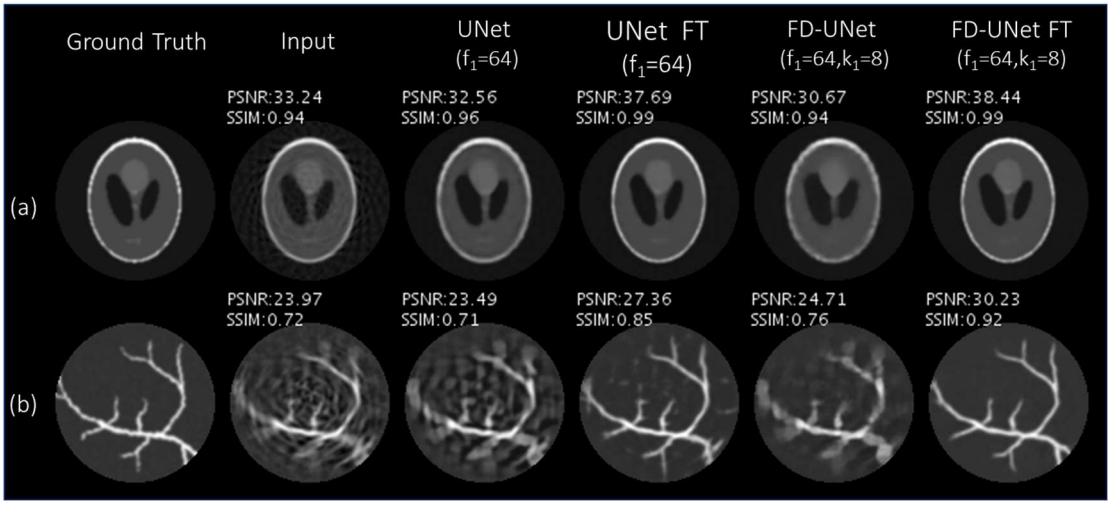

Fig. 14 Reconstructed images (30 sensors) of the (a) Shepp-Logan phantom and (b) vasculature phantom with and without fine-tuning (FT).

Mouse Brain Vasculature Phantom Dataset

As expected, fine-tuning with well-matched training data improves the CNNs' performance, especially in the case of the synthetic vasculature phantom. Both CNNs with fine-tuning recover a sharp and high-quality reconstruction of the Shepp-Logan phantom. Reconstructions of the synthetic vasculature no longer have the circle-like appearance. While both CNNs improve the initial TR reconstruction, the FD-UNet can remove more artifacts and outperform the UNet as evidenced by its higher average PSNR and SSIM for both synthetic phantoms.

In the third experiment, the CNNs are trained on the more complex synthetic vasculature phantom dataset and tested on the mouse brain vasculature dataset. In this scenario, the datasets are well-matched, but there are likely features in the anatomically realistic brain vasculature dataset that are not present in the synthetic vasculature dataset. The purpose of this experiment is to evaluate the feasibility of training the CNNs on

synthetic phantom images for removing artifacts from anatomically realistic vasculature images under multiple levels of sampling sparsity.

Table 5 Average PSNR and SSIM under varying sampling sparsity levels for mouse brain vasculature dataset

# Of Detectors	15	30	45
TR	19.77 ± 0.96 *0.58 ± 0.05*	22.89 ± 1.13 *0.70 ± 0.05*	25.56 ± 1.28 *0.78 ± 0.05*
UNet	20.21 ± 1.19 *0.60 ± 0.07*	22.15 ± 2.35 *0.68 ± 0.11*	25.07 ± 2.09 *0.76 ± 0.11*
FD-UNet	21.12 ± 1.18 *0.65 ± 0.04*	25.13 ± 1.36 *0.82 ± 0.03*	28.47 ± 1.39 *0.89 ± 0.03*

For each row, PSNR is shown as normal text on top while SSIM is shown as italicized text on the bottom. The CNN hyperparameters used are FD-UNet ($f_1 = 64, k_1 = 8$) and UNet ($f_1 = 64$).

As seen in Table 5, there are no significant quantitative changes in PSNR and SSIM between the UNet and TR reconstructions for all levels of sampling sparsity tested. However, the UNet does remove majority of the background artifacts and qualitatively appears better than the TR reconstruction as shown in Fig. 15. No quantitative improvement is observed because the UNet only recovers larger vessels and is missing many of the smaller features. The FD-UNet outperforms the UNet and improves the average PSNR and SSIM. It recovers many of the smaller details that are missing in the UNet reconstruction as shown by the green arrows in Fig. 15b. The performance of the CNNs is heavily dependent on the image quality of the TR reconstruction. Features that are missing in the initial reconstruction are also typically missing or incorrectly reconstructed by the CCNs as shown by the red arrows in Fig. 15a.

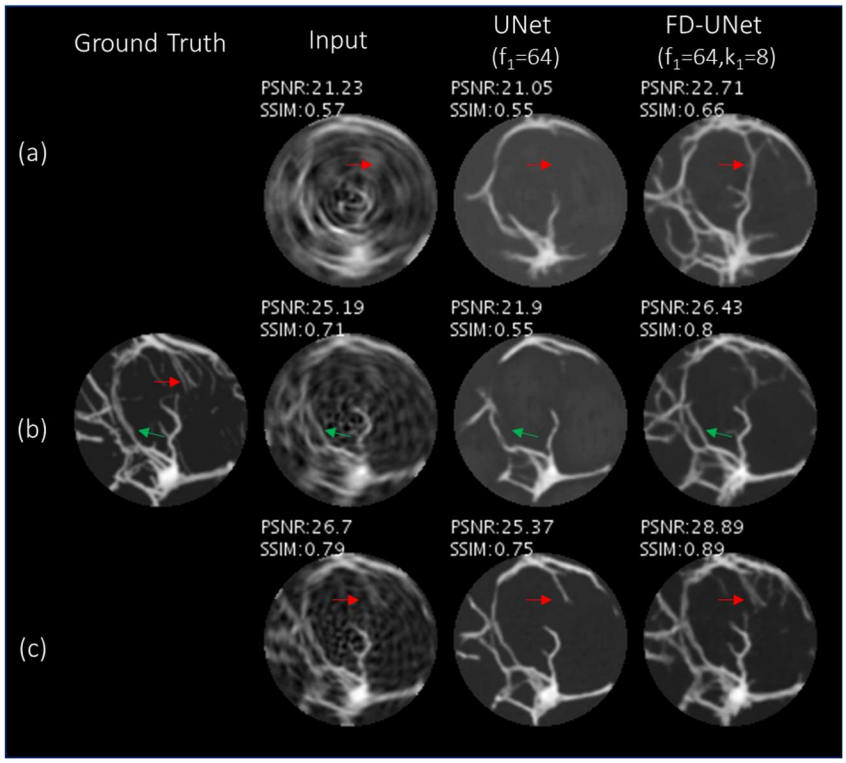

Fig. 15 Examples of reconstructed mouse brain vasculature images for sampling sparsity levels with (a) 15, (b) 30, and (c) 45 detectors. Red and green arrows point to features present in the FD-UNet but missing in the UNet reconstruction.

Section Four: Discussion and Conclusion

In this work, we propose a modified CNN architecture for removing artifacts from 2D PAT images reconstructed from sparse data. Results from the experiments performed consistently show that the FD-UNet is superior to the standard UNet for artifact removal and image enhancement. Dense connectivity strongly encourages feature reuse and improves information flow throughout the network. The benefits in using this connectivity pattern can be observed in Fig. 11. The most compact FD-UNet ($f_1 = 8$) outperforms the more complex UNet ($f_1 = 64$) despite learning fewer features and

requiring only a fraction of the parameters. This demonstrates that the FD-UNet is learning more relevant features for artifact removal, and the ability to reuse those features throughout the network greatly improves the CNN's performance. Furthermore, dense connectivity has a regularizing effect that reduces the likelihood of overfitting to the training data. As seen in Fig. 12., both CNNs converge to a similar PSNR during training yet the FD-UNet outperforms the UNet in testing data. This is likely due to the UNet overfitting to the training data and failing to lean features that generalize well. Furthermore, the UNet training loss is more volatile relative to that of the FD-UNet indicating that the UNet is overfitting to previously observed training examples.

A limitation in using deep learning for artifact removal is that the CNN requires a large training dataset to learn the appropriate weights and features needed to perform well. This limitation can be addressed using computational models (e.g., k-Wave) and synthetic phantoms to generate arbitrarily large datasets for training. However, there remains a challenge in generating a training dataset with all the image features likely to be observed in the testing dataset. This requirement for well-matched training and testing data is demonstrated in the second experiment. As seen in Fig. 14, the CNNs having trained only on images of circles can recover good reconstructions of the Shepp-Logan phantom but not of the synthetic vasculature phantom. Their performance is improved after fine-tuning with a small dataset of synthetic vasculature images. These results provide evidence that it is feasible to initially train the CNN using a poorly matched dataset and then fine-tuned using a small well-matched dataset. This strategy may be useful when only a few relevant experimental training images are available.

In the third experiment, the FD-UNet is trained on the synthetic vasculature dataset and tested on the mouse brain vasculature dataset. While both CNNs remove majority of the background artifacts and reliably recover the larger vessels, the FD-UNet typically recovers more of the smaller vessels than the UNet as seen in Fig. 15. As fewer detectors are used for sampling, the artifacts become increasingly severe in the TR reconstruction and image quality is degraded. A limitation in the post-processing approach is that the CNN's performance strongly depends on the quality of the TR reconstruction. Image features severely obscured by artifacts or missing in the TR reconstruction are likely to be reconstructed incorrectly or missing in the CNN reconstruction. Information is lost as a result of sparse sampling, but the initial step of reconstructing an image from sensor data also discards potentially useful information and introduces artifacts. It may be possible to recover some of the smaller vessels if the CNN is used to directly reconstruct the sensor data into an image.

In this paper, we propose a modified CNN architecture termed FD-UNet for removing artifacts from 2D PAT images reconstructed from sparse data. We compare the FD-UNet and the UNet using datasets generated from synthetic phantoms (circles, Shepp-Logan, and vasculature) and an anatomically realistic mouse brain vasculature dataset. The FD-UNet is demonstrated to be superior and more compact CNN for removing artifacts and improving image quality.

CHAPTER FOUR
PIXEL-WISE DEEP LEARNING

A key limitation in the learned post-processing approach is that the ability of the neural network to remove artifacts is strongly dependent on the quality of the input image. Fine features in the initially reconstructed image are often lost due to artifacts arising from sparse spatial sampling and limited-view detection and are not recoverable by the neural network. We hypothesized that there is potentially useful information in the measured time-series data being lost in the initial image reconstruction step, and this additional information can be leveraged by the neural network to reconstruct a higher quality image. To explore this hypothesis, we developed pixel-wise interpolation as a data pre-processing step that maps the time-series data from the sensor to image space based on the physics of acoustic wave propagation. Pixel-wise interpolation removes the need for an initial image reconstruction and seeks to retain the full information of the time-series data for the neural network to use. By providing a more informative input, the neural network can recover some of the finer features and reconstruct higher quality images. *In Silico* experiments comparing pixel-wise interpolation with other reconstruction methods were performed on several phantoms, and the results were published in Nature Scientific Reports.

Section One: Introduction and Motivation

PAT involves irradiating the biological tissue with a short-pulsed laser. Optical absorbers within the tissue are excited by the laser and undergo thermoelastic expansion

which results in the generation of acoustic waves [11]. A sensor array surrounding the tissue is then used to detect the acoustic waves, and an image is formed from the measured sensor data. PAT image reconstruction is a well-studied inverse problem that can be solved using analytical solutions, numerical methods (e.g. time reversal), and model-based iterative methods [15], [22]–[25]. In general, a high-quality image can be reconstructed if the sensor array has a sufficiently large number of sensor elements and completely encloses the tissue. However, building an imaging system with these specifications is often prohibitively expensive, and in many *in vivo* applications such as neuroimaging, the sensor array typically can only partially enclose the tissue [89], [90]. These practical limitations result in sparse spatial sampling and limited-view of the photoacoustic waves emanating from the medium. Reconstructing from sub-optimally acquired data causes streaking artifacts in the reconstructed PAT image that inhibits image interpretation and quantification [26].

Given the wide success of deep learning in computer vision, there is a strong interest in applying similar methods for tomographic image reconstruction problems [35]–[37]. Deep learning has the potential to be an effective and computationally efficient alternative to state-of-the-art iterative methods. Having such a method would enable improved image quality, real-time PAT image rendering, and more accurate image interpretation and quantification.

Among the many deep learning approaches for image reconstruction, post-processing reconstruction (Post-DL) is the most widely used and has been demonstrated for improving image reconstruction quality in CT [44], [71], MRI [91], and PAT [34],

[46], [61], [63], [74], [92]. It was shown capable of achieving comparable or better performance than iterative methods for limited-view and sparse PAT image reconstruction [45], [61], [93], [94]. In Post-DL, an initial inversion is used to reconstruct an image with artifacts from the sensor data. A convolutional neural network (CNN) is then applied as a post-processing step to remove artifacts and improve image quality. The main drawback of Post-DL is that the initial inversion does not properly address the issues of limited-view and sparse sampling, which results in an initial image with artifacts. Image features (e.g., small vessels) that are missing or obscured by artifacts are unlikely to be recovered by the CNN.

Previous works attempted to improve upon Post-DL by removing the need for an initial inversion step [45], [59]. One approach termed direct reconstruction (Direct-DL) used a CNN to reconstruct an image directly from the sensor data [59]. The main challenge in using Direct-DL is the need to carefully select parameters (e.g., stride and kernel size) for each convolutional layer in order to transform the sensor data into the desired image dimensions. Changing either the dimensions of the input (e.g., using a different number of sensors) or output would require a new set of convolution parameters and the CNN architecture to be modified. Direct-DL was shown capable of reconstructing an image but underperformed compared to Post-DL. Interestingly, a hybrid approach using a combination of Post-DL and Direct-DL, where an initial inversion and the sensor data are given as inputs to the CNN, was shown to provide an improvement over using Post-DL alone [95], [96].

Another approach termed "model-based learning" similarly does not require an initial inversion step and achieves state-of-the-art image reconstruction quality [45], [97]–[99]. This approach is like iterative reconstruction and uses an explicit model of photoacoustic wave propagation for image reconstruction. However, the prior constraints are not handcrafted and instead are learned by a CNN from training data. The improved performance does come at the cost of requiring more time to train the CNN and reconstruct an image [45]. Thus, the choice between model-based learning and direct learned approaches (e.g., Post-DL and Direct-DL) depends on whether the application prioritizes image reconstruction speed or quality.

In this work, we propose a novel approach termed pixel-wise deep learning (Pixel-DL) for limited-view and sparse PAT image reconstruction. Pixel-DL is a direct learned approach that employs pixel-wise interpolation to window relevant information, based on the physics of photoacoustic wave propagation, from the sensor data on a pixel-basis. The pixel-interpolated data is provided as an input to the CNN for image reconstruction. This strategy removes the need for an initial inversion and enables the CNN to utilize more information from the sensor data to reconstruct a higher quality image. The pixel-interpolated data has similar dimensions to the desired output image which simplifies CNN implementation. We compare Pixel-DL to conventional PAT image reconstruction methods (time reversal and iterative reconstruction) and direct learned approaches (Post-DL and a modified implementation of Direct-DL) with *in silico* experiments using several vasculature phantoms for training and testing.

Section Two: Methods

In this work, three different CNN-based deep learning approaches were used for limited-view and sparse PAT image reconstruction (Fig. 16). These direct learned approaches all began with applying an initial processing step to the PAT sensor data and then recovering the final PAT image using a CNN. The primary difference among these approaches was the processing step used to initially transform the PAT sensor data. In Post-DL, the sensor data was initially reconstructed into an image containing artifacts using time reversal, and the CNN was applied as a post-processing step for artifact removal and image enhancement. In Pixel-DL, pixel-wise interpolation was applied to window relevant information in the sensor data and to map that information into the image space. In the modified Direct-DL implementation (mDirect-DL), a combination of linear interpolation and down sampling was applied so that the interpolated sensor data had the same dimensions as the final PAT image.

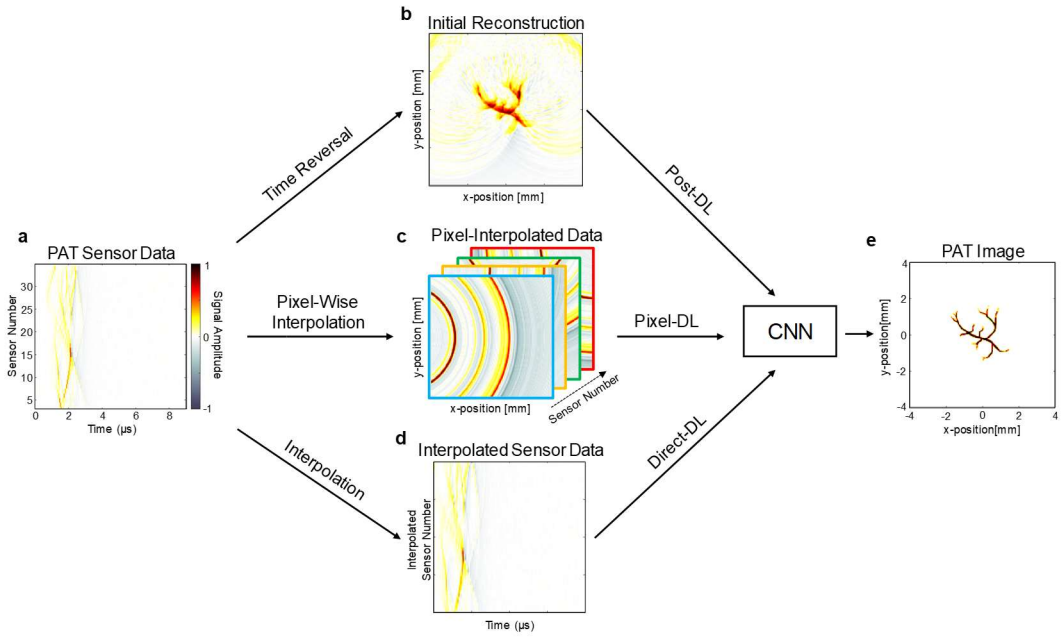

Fig. 16 Summary of CNN-based deep learning approaches for PAT image reconstruction. The primary task is to reconstruct an essentially artifact-free PAT image from the acquired PAT sensor data. a) PAT sensor data acquired using a sensor array with 32 sensors and semi-circle limited-view. b) Initial image reconstruction with sparse and limited-view artifacts using time reversal for Post-DL. c) 3D data array acquired after applying pixel-wise interpolation for Pixel-DL. d) Sensor data interpolated to have matching dimensions as the final PAT image for mDirect-DL. e, Desired artifact-free PAT image reconstruction from the CNN-based deep learning approaches.

After the sensor data was transformed, the final PAT image was recovered using the Fully Dense UNet (FD-UNet) CNN architecture (Fig. 17). The FD-UNet builds upon the UNet, a widely used CNN for biomedical imaging tasks. by incorporating dense connectivity into the contracting and expanding paths of the network [66]. This connectivity pattern enhances information flow between convolutional layers to mitigate learning redundant features and reduce overfitting [65]. The FD-UNet was demonstrated to be superior to the UNet for artifact removal and image enhancement in 2D sparse PAT [63].

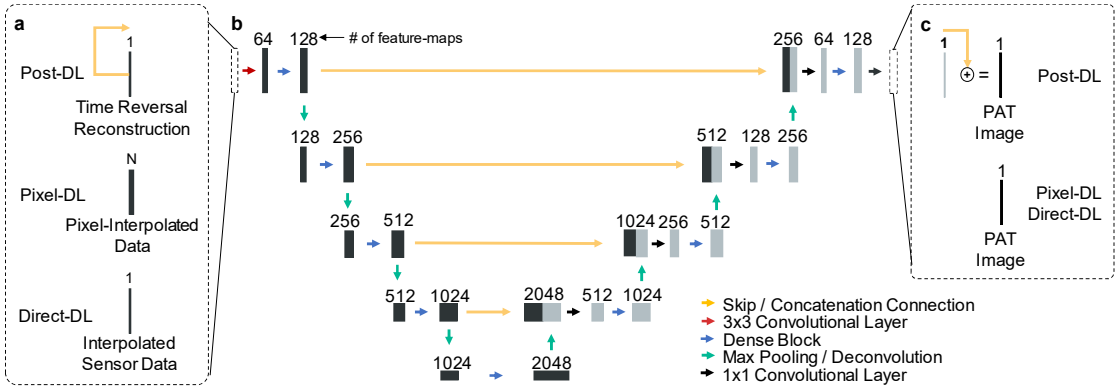

Fig. 17 FD-UNet CNN Architecture. The FD-UNet CNN with hyperparameters of initial growth rate, $k_1 = 16$ and initial feature-maps learned, $f_1 = 128$ is used for PAT image reconstruction. Essentially the same CNN architecture was used for each deep learning approach except for minor modifications. a) Inputs into the CNN for each deep learning approach. The Post-DL CNN implementation used residual learning which included a skip connection between the input and final addition operation. The initial Pixel-DL input contains "N" feature-maps corresponding to the number of sensors in the imaging system. b) The FD-UNet is comprised of a contracting and expanding path with concatenation connections. c) The output of the CNN is the desired PAT image. In Post-DL, residual learning is used to acquire the final PAT image.

Pixel-wise Interpolation

Pixel-wise interpolation uses a model of photoacoustic wave propagation to map the measured time series pressure in the sensor data to a pixel position within the image reconstruction grid that the signal likely originated from. In this work, we choose to apply pixel-wise interpolation using a linear model of photoacoustic wave propagation since the *in silico* experiments were performed using a homogenous medium (e.g. uniform density and speed of sound). The linear model assumes the acoustic waves are propagating spherically and traveling at a constant speed of sound. Based on these assumptions, the time-of-flight can be easily calculated for a pressure source originating at some position in the medium and traveling to a sensor located on the medium boundary.

Reconstructing an image begins by defining an image reconstruction grid that spans the region of interest in the imaging system (Fig. 18a). The goal of pixel-wise interpolation is to map the time series pressure measurements of each sensor to the defined reconstruction grid on a pixel-basis, which results in a 3D data array with dimensions corresponding to the 2D image space and sensor number (Fig. 18b-c). This is achieved by repeating the following interpolation process for each sensor in the sensor array (Fig. 18d-f). The time-of-flight for a signal originating from each pixel position and traveling to the selected sensor is calculated based on a model of photoacoustic wave propagation. In the case of a linear model, the time-of-flight is proportional to the distance between the selected pixel and sensor (Fig. 18e). Pressure measurements in the sensor data are interpolated onto the reconstruction grid using the calculated time-of-flight for each pixel (Fig. 18f).

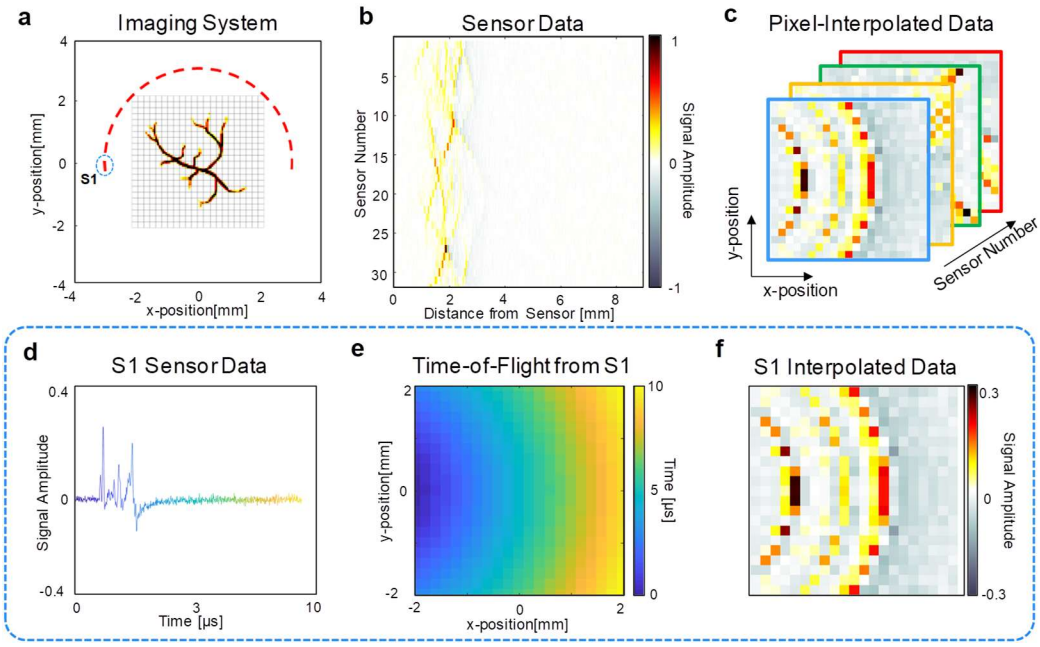

Fig. 18 Pixel-Wise Interpolation Process. a) Schematic of the PAT system for imaging the vasculature phantom. The red semi-circle represents the sensor array, and the gray grid represents the defined reconstruction grid. The first sensor (S1) is circled and used as an example for applying pixel-wise interpolation to a single sensor. b) The PAT time series pressure sensor data measured by the sensor array. c) Resulting pixel-interpolated data after applying pixel-wise interpolation to each sensor based on the reconstruction grid. d) Sensor data for S1. Color represents the time at which a pressure measurement was taken and is included to highlight the use of time-of-flight to map the sensor data to the reconstruction grid. e) Calculated time-of-flight for a signal originating at each pixel position and traveling to S1. f) Pressure measurements are mapped from the S1 sensor data to the reconstruction grid based on the calculate time-of-flight for each pixel.

Deep Learning Implementation

The CNNs were implemented in Python 3.6 with TensorFlow v1.7, an open source library for deep learning [87]. Training and evaluation of the network is performed on a GTX 1080Ti NVIDIA GPU. The CNNs were trained using the Adam optimizer to minimize the mean squared error loss with an initial learning rate of 1e-4 and a batch size of three images for 40 epochs. Training each CNN required approximately one hour to complete. Pairs of training datasets $\{x_i, y_i\}$ were provided to the CNN during training, where x_i represents the input data (e.g., initial time reversal reconstruction, pixel-

interpolated sensor data, and interpolated sensor data) and y_i represents the corresponding artifact-free ground truth image. A separate CNN was trained for each CNN-based approached, imaging system configuration, and training dataset.

Photoacoustic Data for Training and Testing

Training data were procedurally generated using data augmentation, where new images were created based on a 340x340 pixel-size image of a synthetic vasculature phantom generated in MATLAB (Fig. 18a). First, scaling and rotation was applied to the initial phantom image with a randomly chosen scaling factor (0.5 to 2) and rotation angle (0-359 degrees). Then a 128x128 pixels sub-image was randomly chosen from the transformed image and translated by a random vertical and horizontal shift (0-10 pixels) via zero-padding. Outputs from multiple iterations (up to five) of the data augmentation process are summed together to create a training image. The synthetic vasculature phantom dataset was comprised of 500 training images. Testing data were generated from a 3D micro-CT mouse brain vasculature volume [85] with a size of 260x336x438 pixels. The Frangi vesselness filter was applied to suppress background noise and enhance vessel-like features [86]. A new image was created from the filtered volume by generating a maximum-intensity projection of a randomly chosen 128x128x128 pixel sub-volume. The mouse brain vasculature dataset was comprised of 50 testing images.

The “High-Resolution Fundus Image Database” is a public database that contains 45 fundus images from human subjects that were either healthy, had glaucoma, or had diabetic retinopathy. The images had corresponding vessel segmentation maps created by

a group of experts and clinicians within the field of retinal image analysis [100]. The 45 fundus images were split into a separate training set (N=15) and testing set (N=30). The training dataset was procedurally generated using data augmentation based on the images within the training set and was comprised of 500 training images. The testing dataset was comprised of the original 30 images and 20 additional images, generated using data augmentation based on images from the testing set, for a total of 50 testing images.

The "ELCAP Public Lung Image Database" is a public database that contains 50 low-dose whole-lung CT scans obtained within a single breath hold [101]. The whole-lung volumes were split into a training (N=15) and testing set (N=35). Vessel-like structures were segmented from the whole-lung CT volumes using the Frangi vesselness filter [63]. The training dataset was then generated by taking maximum intensity projection images (MIP) of randomly sampled sub-volumes from the filtered volumes in the training set. Data augmentation was also applied to the MIPs to generate a training dataset comprised of 500 training images. With the same procedures, MIPs were taken from the filtered volumes in the testing set to create a testing dataset comprised of 50 images.

In all three cases (mouse-brain vasculature, fundus image database, and ELCAP Lung database), training and testing data were completely segregated. In the latter two experiments, significant variations were present between the training and testing datasets due to patient-to-patient variability and innate differences in vascular morphology between healthy subjects and patients with varying degrees of disease.

A MATLAB toolbox, k-WAVE, was used to simulate photoacoustic data acquisition using an array of acoustic sensors [83]. Photoacoustic simulations in the k-WAVE toolbox are implemented using a pseudospectral approach [102]. Each training and testing image were normalized (values between 0 and 1) and treated as a photoacoustic source distribution on a computation grid of 128x128 pixels. The medium was assumed to be non-absorbing and homogenous with a speed of sound of 1500 m/s and density of 1000 Kg/m^3. The sensor array had 16, 32, or 64 sensor elements equally spaced on a semi-circle with a diameter of 120 pixels. The time reversal method in the k-WAVE toolbox was also used for reconstructing an image from the simulated photoacoustic time series data.

Reconstructed images were compared against the ground truth using the peak-signal-to-noise ratio (PSNR) and structural similarity index (SSIM) as metrics for image quality. PSNR provides a global measurement of image quality, while SSIM provides a local measurement that takes into account for similarities in contrast, luminance, and structure [88].

Section Three: Methods

Conventional PAT image reconstruction techniques (e.g., time reversal and iterative reconstruction) and CNN-based approaches (Post-DL, Pixel-DL, and mDirect-DL) were compared over several *in silico* experiments for reconstruction image quality and reconstruction time. CNN-based approaches were all implemented using the FD-

UNet CNN architecture. Reconstructed images were compared to the ground truth image using PSNR and SSIM as quantitative metrics for image reconstruction quality.

Mouse Brain Vasculature Experiment

In the first experiment, the CNNs were trained on the synthetic vasculature phantom dataset and tested on the mouse brain vasculature dataset. Although both datasets contained images of vasculature, they were non-matched meaning there were likely image features (e.g., vessel connectivity patterns) in the testing dataset but not in the training dataset. In addition to evaluating the CNNs' performance, this experiment sought to determine if the CNNs were generalizable when trained on the synthetic vasculature phantom and tested on the mouse brain datasets.

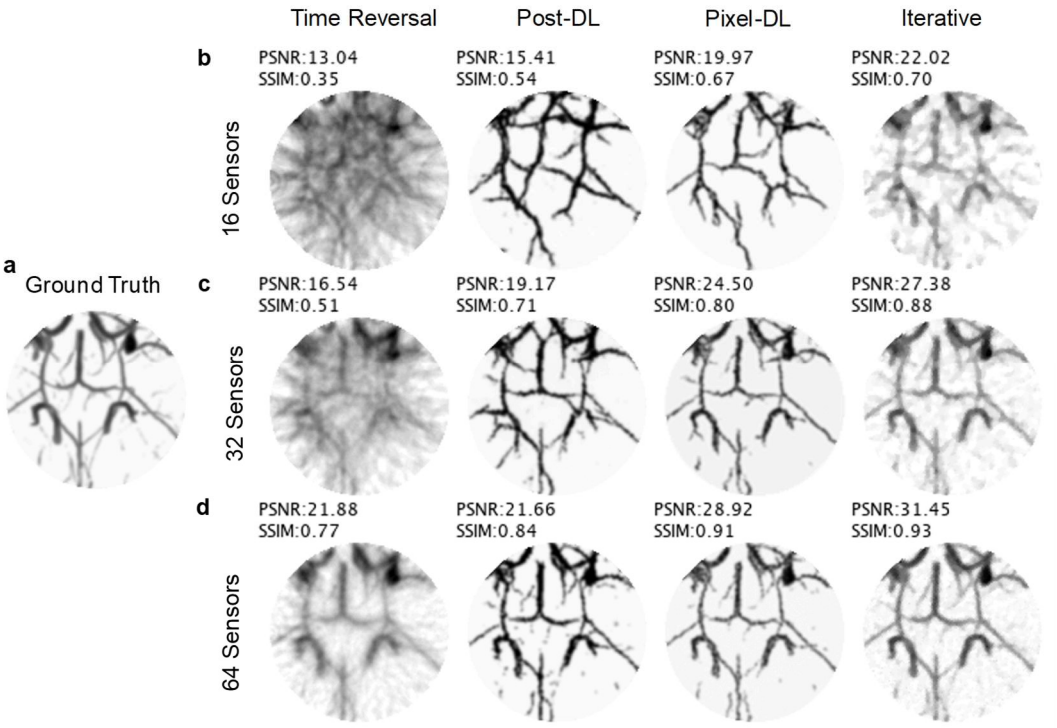

Fig. 19 Limited-view and sparse PAT image reconstruction of mouse brain vasculature. PAT sensor data acquired with a semi-circle limited-view sensor array at varying sparsity levels. a) Ground truth image used to simulate PAT sensor data. b) PAT reconstructions with 16 sensors. Vessels are difficult to identify in time reversal reconstruction as a result of artifacts. c) PAT reconstructions with 32 sensors. Vessels can be clearly seen in CNN-based and iterative reconstructions. d) PAT reconstructions with 64 sensors. Larger vessels are identifiable in all reconstructed images.

The time reversal reconstructed images had severe artifacts blurring the image and the lowest average PSNR and SSIM for all sparsity levels (Fig. 19 and Table 6). Images reconstructed with iterative or a CNN-based method had fewer artifacts and a higher average PSNR and SSIM. Vessels obscured by artifacts in the time reversal reconstructed images were more visible in the other reconstructed images. As expected, increasing the number of sensors resulted in fewer artifacts and improved image quality for all PAT image reconstruction methods. Pixel-DL consistently had a higher average

PSNR and SSIM than Post-DL for all sparsity levels and similar scores to iterative reconstruction.

Table 6 Average PSNR and SSIM for Micro-CT Mouse Brain Vasculature Testing Dataset (N = 50)

Number of Sensors	Time Reversal	Post-DL	Pixel-DL	Iterative Reconstruction
16	13.91±1.12 *0.34±0.04*	17.4±1.24 *0.52±0.04*	21.52±1.36 *0.64±0.04*	22.64±1.4 *0.66±0.05*
32	17.29±1.20 *0.48±0.04*	21.31±1.10 *0.71±0.04*	25.67±1.29 *0.81±0.04*	26.98±2.11 *0.82±0.06*
64	22.7±1.06 *0.73±0.03*	24.37±1.25 *0.85±0.03*	29.59±1.42 *0.91±0.02*	30.16±2.70 *0.89±0.05*

For each row, PSNR is shown as normal text on top while SSIM is shown as italicized text on the bottom.

In the case of sparse sampling (especially with 16 sensors), Post-DL often introduced additional vessels that were not originally in the ground truth image (Fig. 19a-b). This was likely due to the CNN misinterpreting strong artifacts in the input image as real vessels. Pixel-DL exhibited a similar behavior but typically had fewer false additional vessels. This issue was not as prevalent in images reconstructed using the iterative method. However, images reconstructed using iterative reconstruction had an overly smoothed appearance compared to the deep learning-based reconstructed images. This is a pattern commonly observed when using the total variation constraint.

Pixel-DL consistently outperformed time reversal in reconstructing images of the synthetic vasculature and mouse brain vasculature (Fig. 20). Interestingly, mDirect-DL only outperformed time reversal in reconstructing the synthetic vasculature images, which were used to train the CNN. The mDirect-DL reconstructed image of mouse brain

vasculature resembled the ground truth image but was substantially worse than the time reversal reconstruction. This indicated that the CNN learned a mapping from the PAT-sensor data to the image space but severely overfitted to the training data. During training, the CNNs for Pixel-DL and mDirect-DL converged to a minimum mean squared error, but the Pixel-DL CNN converged to a lower error.

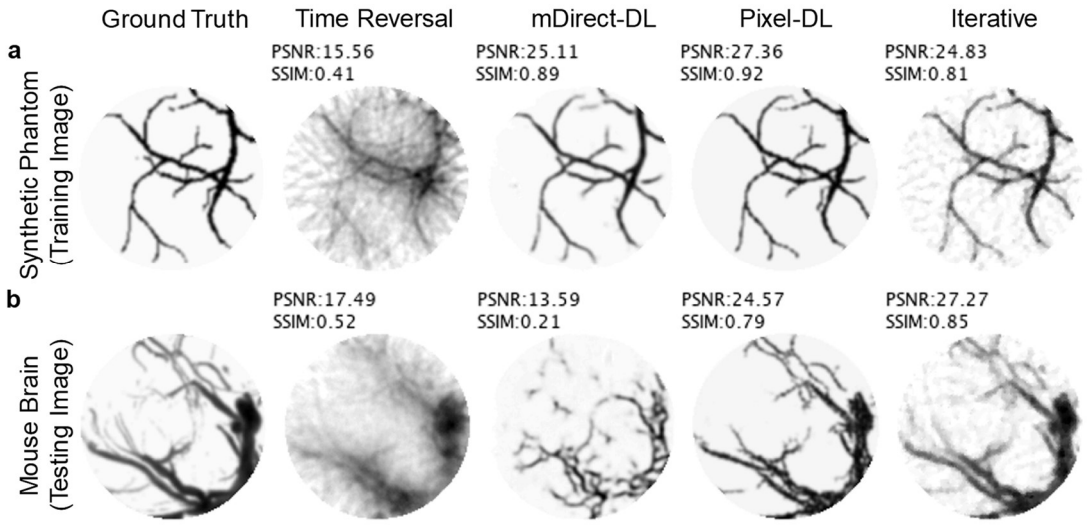

Fig. 20 Limited-view and sparse Pixel-DL and mDirect-DL PAT image reconstructions. PAT sensor data acquired with 32 sensors and a semi-circle view. a) CNNs were trained and tested on images of the synthetic vasculature phantom. Both CNN-based approaches successfully reconstructed the example synthetic vasculature phantom image b) CNNs were trained on images of the synthetic vasculature phantom but tested on mouse brain vasculature images. mDirect-DL failed to reconstruct the example mouse brain vasculature image and performed worse than time reversal.

Lung and Fundus Vasculature Experiment

In the second experiment, the CNNs were trained and tested on the lung vasculature and fundus vasculature datasets. This experiment represented a scenario in

which the training and testing datasets are derived from segregated anatomical image data. There were natural differences between the training and testing datasets since the original images were acquired from healthy patients and those with varying disease severity.

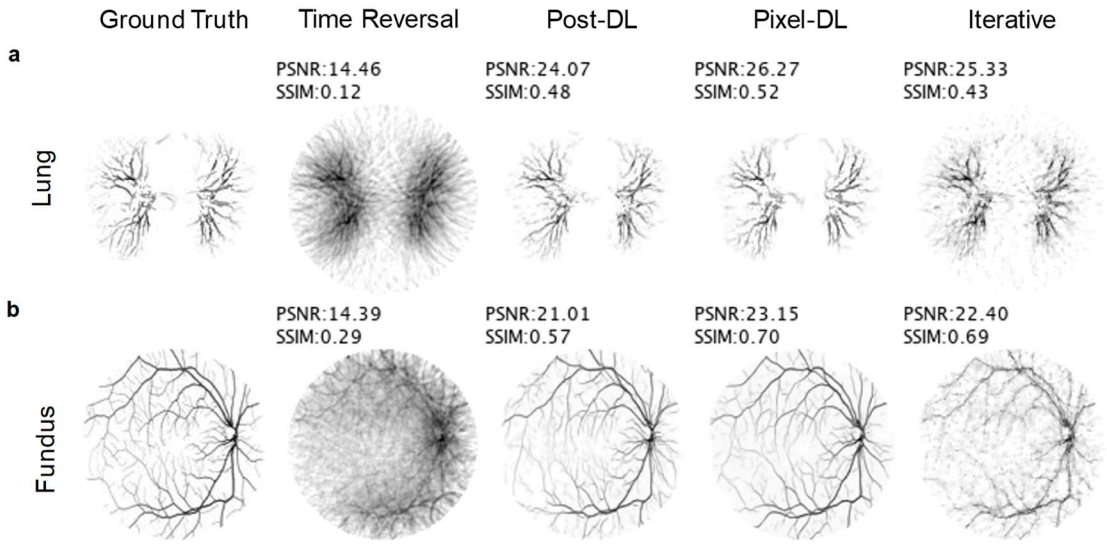

Fig. 21 Limited-view and sparse PAT image reconstructions of fundus and lung vasculature. PAT sensor data acquired with 32 sensors and a semi-circle view. a) CNNs were trained and tested on images of lung vasculature b) CNNs were trained and tested on images of fundus vasculature. Testing images were derived from a separate set of patients' lung and fundus images than the training images.

As expected, the time reversal reconstructed images of lung and fundus vasculature had the most artifacts and the lowest average PSNR and SSIM for all sparsity levels (Fig. 21 and Table 7). Images reconstructed with a CNN-based method or iterative reconstruction resulted in fewer artifacts and a higher average PSNR and SSIM. Pixel-DL consistently outperformed Post-DL for both vasculature phantoms for all sparsity levels. Comparable to iterative reconstruction, Pixel-DL had similar performance for the fundus

vasculature and outperformed it for the lung vasculature dataset. For images reconstructed from PAT sensor data acquired using 16 sensors, Pixel-DL reconstructed images appeared sharper and were qualitatively superior compared to iteratively reconstructed images despite having similar SSIM and PSNR values.

Table 7 Average PSNR and SSIM for Lung and Fundus Vasculature Testing Dataset (N = 50 testing images)

	Number of Sensors	Time Reversal	Post-DL	Pixel-DL	Iterative Reconstruction
Lung	16	13.30±1.01 *0.09±0.02*	23.21±1.45 *0.35±0.04*	24.14±1.53 *0.43±0.06*	22.74±1.36 *0.29±0.08*
	32	15.19±1.13 *0.13±0.02*	25.09±1.67 *0.50±0.04*	26.76±1.83 *0.53±0.07*	27.50±1.98 *0.46±0.06*
	64	18.82±1.11 *0.23±0.05*	27.14±1.67 *0.65±0.04*	29.98±2.00 *0.69±0.11*	33.67±1.92 *0.62±0.07*
Fundus	16	12.26±1.10 *0.19±0.02*	20.00±1.52 *0.42±0.06*	20.78±1.61 *0.52±0.08*	20.77±1.07 *0.50±0.04*
	32	14.07±1.38 *0.26±0.03*	21.57±1.60 *0.59±0.04*	23.40±1.40 *0.67±0.05*	23.37±1.06 *0.68±0.04*
	64	18.08±1.40 *0.45±0.05*	24.16±1.56 *0.75±0.03*	26.23±1.35 *0.81±0.05*	28.07±1.10 *0.85±0.06*

For each row, PSNR is shown as normal text on top while SSIM is shown as italicized text on the bottom.

Image Reconstruction Times

The average reconstruction time reported for each method are for reconstructing a single image from the PAT sensor data. Time reversal is a robust and computationally inexpensive reconstruction method (~2.57 seconds per image). Iterative reconstruction removed most artifacts and improved image quality but had a much longer average reconstruction time (~491.21 seconds per image). Pixel-DL reconstructed images with similar quality to iterative reconstruction and was faster by over a factor of 1000 (~7.9 milliseconds per image). Average reconstruction time for Post-DL is dependent on the

initial inversion used since the computational cost of a forward pass through a CNN is essentially negligible. Since time reversal was used as the initial inversion, Post-DL had a longer average reconstruct time than Pixel-DL (~2.58 seconds per image).

Section Four: Discussion and Conclusion

In this work, we propose a novel deep learning approach termed Pixel-DL for limited-view and sparse PAT image reconstruction. We performed *in silico* experiments using training and testing data derived from multiple vasculature phantoms to compare Pixel-DL with conventional PAT image reconstruction methods (time reversal and iterative reconstruction) and direct learned approaches (Post-DL and mDirect-DL). Results showed that Pixel-DL consistently outperformed time reversal, Post-DL, and mDirect-DL for all experiments. Pixel-DL was able to generalize well evidenced by its comparable performance to iterative reconstruction for the mouse brain vasculature phantom despite having only trained on images generated from a synthetic vasculature phantom with data augmentation. Having a more varied training dataset may further improve CNN generalization and performance. When the training and testing data were derived from segregated anatomical data, Pixel-DL had similar performance to iterative reconstruction for the fundus vasculature phantom and outperformed it for the lung vasculature phantom. The total variation constraint used for iterative reconstruction was likely suboptimal for reconstructing lung vasculature images since the lung vessels were small and closely grouped.

Comparison between Deep Learning Frameworks

The CNN architecture and hyperparameters used for all deep learning approaches implemented were essentially the same. Thus, discrepancies in performance between the approaches were primarily due to their respective inputs into the CNN. In Post-DL, the input was an image initially reconstructed from the sensor data using time reversal. The input and output to the CNN are both conveniently images of the same dimensions. This removed the need for the CNN to learn the physics required to map the sensor data into the image space. However, the initial inversion did not properly address the issues of limited-view and sparse sampling which resulted in an initial image with artifacts. Moreover, the CNN no longer had access to the sensor data and was only able to use information contained in the image to remove artifacts. There was likely useful information in the sensor data for more accurately reconstructing the PAT image, which was ignored in this approach.

In Pixel-DL, the initial inversion is replaced with pixel-wise interpolation, which similarly provides a mapping from the sensor data to image space. Relevant sensor data is windowed on a pixel-basis using a linear model of acoustic wave propagation. This enables the CNN to have a richer information source to reconstruct higher quality images. Furthermore, there is no initial inversion introducing artifacts; thus, the CNN does not have an additional task of learning to remove those artifacts.

mDirect-DL similarly did not require an initial inversion and instead used the full sensor data as an input to the CNN to reconstruct an image. The potential advantage of mDirect-DL is that the CNN had full access to the information available in the sensor

data to reconstruct a high-quality image. However, reconstructing directly from the sensor data was also a more difficult task because the CNN needed to additionally learn a mapping from the sensor data into the image space. Results showed that the CNN had difficulty in learning a generalizable mapping and overfitted to the training data. The FD-UNet was likely not an optimal architecture for this task since it was designed assuming the input was an image. A different neural network architecture for a multidimensional time-series input would be better suited.

A limitation of Post-DL and Pixel-DL for sparse and limited-view PAT is that the reconstructed image could have additional vessels that are not in the ground truth image. This can be problematic depending on the requirements of the application. Large vessels and structures are often reliably reconstructed in the image, but some small vessels could be false additions. This limitation primarily occurred at the sparsest sampling level and could be addressed by increasing the number of sensors used for imaging. The loss function could also be modified to penalize the CNN for reconstructing false additional vessels, but this could lead to the CNN to preferentially not reconstruct small vessels. Alternatively, a model-based learning approach could be used for better image quality if computational cost is not a limitation.

Deep Learning for In Vivo Imaging

A key challenge in applying deep learning for *in vivo* PAT image reconstruction is that a large training dataset is required for the CNN to learn and be able to remove artifacts and improve image quality. The training data can be acquired experimentally

using a PAT imaging system that has a sufficient number of sensors and full view of the imaging target. However, this process is often infeasible because it is prohibitively expensive, time-consuming, and needs to be repeated when the imaging system configuration or imaging target is changed. Alternatively, synthetic training data can be generated using numerical phantoms or images from other modalities. In combination with data augmentation techniques, this approach enables for arbitrarily large synthetic training datasets to be created. However, CNN image reconstruction quality is largely dependent on the degree to which the simulations used to generate the training data matches actual experimental conditions. Properly matching the simulation is a non-trivial task that necessitates the PAT imaging system to be well-characterized and understood. Some factors to be considered when creating the simulations include sensor properties (e.g., aperture size, sensitivity, and directivity), sensor configuration, laser illumination, and medium heterogeneities. Generally, it is preferable to closely match the simulation to the experimental conditions, but post-processing (e.g., filtering and denoising) can also be applied to the experimental data. It is beyond the scope of this work to discuss the impact of each factor in detail, but the issue of medium heterogeneities, specifically for speed of sound, is examined.

In this work, Pixel-DL was applied using a linear model of acoustic wave propagation that assumes the acoustic waves propagate spherically and travel at a constant speed of sound throughout the medium. Although this model was sufficient for the case of a homogenous medium, a different model would be needed if the medium was heterogeneous (e.g., speed of sound and density) such as for *in vivo* imaging. Naively

reconstructing with these assumptions for heterogeneous mediums would result in additional artifacts that degrade image quality and potentially impact CNN performance. The severity of the artifacts would depend on the degree of mismatch between the heterogeneity and assumed value. If the distribution of the heterogeneities or acoustically reflective surfaces is known, then they can be accounted for during the time-of-flight calculations when applying pixel-interpolation. However, if it is not known then the CNN should be trained with training data containing examples of heterogeneous mediums like what would be anticipated during image reconstruction. This would enable the CNN to learn to compensate for potential artifacts due to applying pixel interpolation with a linear model of acoustic wave propagation when the medium is not homogeneous.

Deep Learning for Fast Image Reconstruction

The proposed Pixel-DL approach can be used as a computationally efficient method for improving PAT image quality under limited-view and sparse sampling conditions. It can be readily applied to a wide variety of PAT imaging applications and configurations. Pixel-DL enables for the development of more efficient data acquisition approaches. For example, PAT imaging systems can be built with fewer sensors without sacrificing image quality, which would allow for the technology to be more affordable. Pixel-DL achieved similar or better performance and was faster than iterative reconstruction by over a factor of a 1000. It would allow for real-time PAT image rendering which would provide valuable feedback during image acquisition.

In this work we have demonstrated *in silico* the feasibility of Pixel-DL for PAT imaging of vasculature-like targets. This approach can also be readily applied to ultrasound imaging. Image reconstruction for PAT and ultrasound imaging both largely rely on time-of-flight calculations to determine where the signal originated. Therefore, a similar linear model of acoustic wave propagation can be used to readily apply Pixel-DL for ultrasound image reconstruction problems. Pixel-DL can also be adapted to other imaging modalities if a model mapping the sensor data to the image space is available.

CHAPTER FIVE
DENSE DIALTED UNET FOR 3D PAT

A key challenge in 3D PAT image reconstruction is the large memory requirement and computational cost in manipulating a multi-channel 4D array. These limitations prevent the use of more complex CNN architectures with many layers and learnable parameters. While simpler CNNs can be used for 3D PAT image reconstruction, the CNN may not perform well due to its limited complexity and not learning the necessary features for its defined task. In this work, we sought to improve upon the widely used UNet CNN architecture by incorporating dense connectivity and dilated convolutions into the network structure. These modifications would allow the CNN to learn more meaningful and useful features for reconstructing a high-quality image without increasing CNN model complexity.

Section One: Introduction and Motivation

Many deep learning approaches have been developed for PAT image reconstruction [58], [103], [104]. Post-processing reconstruction (Post-DL) is the most widely used and has been previously demonstrated for removing artifacts and improving image quality in PAT and other imaging modalities such as CT and MRI [61], [105], [106]. In Post-DL, an initial image with artifacts is reconstructed from the time-series data, and a convolutional neural network (CNN) is applied as a post-processing step to remove artifacts [61], [63]. The main drawback of Post-DL is that potentially useful information in the time-series data is lost during the initial inversion. Other approaches

(e.g. *Pixel-DL* and the *upgUNET)* improve upon Post-DL by replacing the initial inversion with a data pre-processing step to provide a more informative input for the CNN [107], [108]. Direct-learned reconstructions seeks to reconstruct an image directly from the time-series data with a CNN but often underperform compared to Post-DL [103]. Among the different approaches, model-based reconstruction was shown to outperform other deep learning approaches [45]. Like iterative reconstruction, this approach uses an explicit model of photoacoustic wave propagation, but the prior constraints are instead learned from data. The improved performance comes at the cost of increased computational complexity and slower image reconstruction.

In this work, the Post-DL approach is followed for 3D PAT reconstruction of sparse imaging targets in applications requiring fast image reconstruction. Although Pixel-DL has been shown to outperform Post-DL in 2D PAT, it is not suitable for 3D PAT imaging due to the large memory requirement and computational cost for manipulating the 4D pre-processed data array. We propose a modified CNN architecture termed Dense Dilated UNet (DD-Net) for 3D PAT imaging of sparse targets in a heterogeneous medium. This work builds upon the well-known UNet CNN architecture for biomedical imaging by incorporating dense connectivity and dilated convolutions throughout the network. Dense connectivity enables the CNN to learn more diverse feature sets by mitigating the need to relearn redundant features and enhancing information flow [65]. Dilated convolutions expand the CNN's effective receptive field without loss of resolution or coverage for learning multi-scale context [109].

Section Two: Methods

An image containing streaking artifacts is initially reconstructed from the incomplete time-series data using time reversal (Fig. 22c). A CNN is then applied as a post-processing step to remove artifacts and improve image quality (Fig. 22d). This task can be formulated as a supervised learning problem, in which the CNN learns a function that maps the input, an image with artifacts, to the desired output, an artifact-free image [64]. The CNN is trained on paired examples of the initial time reversal reconstruction and the ground truth image.

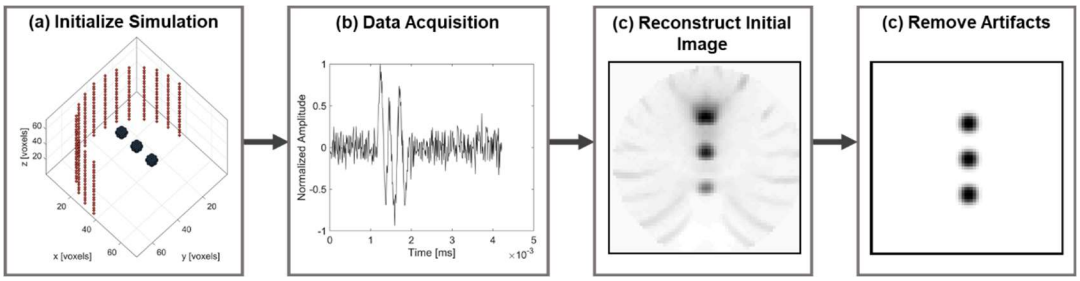

Fig. 22 Process diagram demonstrating the generation of sparse spatial sampling and limited-view 3D PAT data and Post-DL image reconstruction. (a) Simulation was initialized using a cylindrical sensor configuration (red elements) with a half-circle view and sparse spatial sampling to image spherical objects (black elements) in the center. (b) Example time-series data for a single sensor element with added Gaussian noise (25 dB PSNR). (c) Maximum intensity projection through the z-axis of the 3D image with artifacts when reconstructed using the time reversal method. (d) Maximum intensity projection through the z-axis of the 3D image without artifacts after post-processing using a CNN.

Dilated Convolutions

The dilated convolution, also known as the atrous convolution, is an extension of the standard convolution, in which the convolutional filter is upsampled by inserting

zeros between the weights [110]. In the 1-D case of a dilated convolution, the output $\boldsymbol{o}$ at location $\boldsymbol{i}$ with a filter $\boldsymbol{w}$ of size $\boldsymbol{S}$ and dilation rate $\boldsymbol{r}$ for an input $\boldsymbol{f}$, can be represented as

$$o[i] = \sum_{s=1}^{S} f[i + r \cdot s]w[i] \qquad (1)$$

When the dilation rate is one, the dilated convolution is equivalent to a standard convolution. A key advantage in using dilated convolutions is that the receptive field of the convolution operation can be enlarged without requiring additional training parameters (Fig. 23). The receptive field describes the area of an image that can be viewed by an artificial neuron to extract information. A larger receptive field is needed to learn multi-scale features which is conventionally achieved by connecting successive convolutional layers in a cascade and using max pooling layers to spatially down sample the image [110]. Dilated convolutions allow the CNN to more efficiently learn multi-scale features without a rescaled image and loss of resolution. Cascaded dilated convolutions also expand the receptive field exponentially, whereas, cascaded standard convolutions expands it linearly [111].

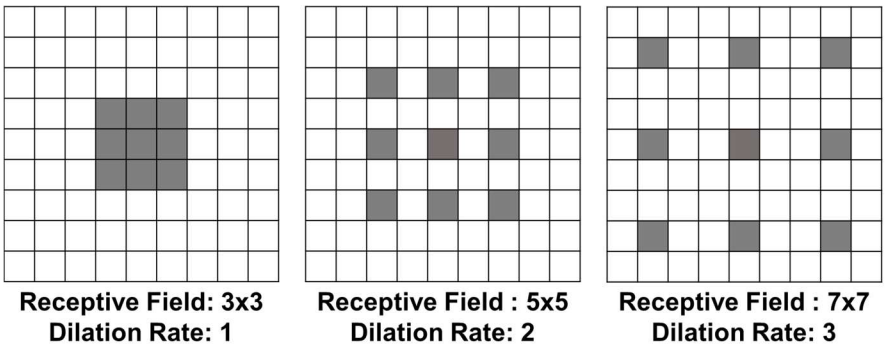

Fig. 23. In a dilated convolution, the effective receptive field of the convolution operation is enlarged by inserting gaps between the kernel weights of a 3x3 filter based on the dilation rate.

However, it has been observed that the use of dilated convolutions results in "gridding artifacts" [110], [112]. Because of the zero-padded gaps in the convolutional filter, adjacent units in the output are calculated from completely separate inputs. Therefore, gridding artifacts occur when the image or feature map has higher-frequency content than the sampling rate of the dilated convolution [109]. Artifacts tend to be more severe for larger dilation rates and with cascaded dilated convolutions.

Dense Dilation Blocks

For increasingly complex tasks, a deeper CNN with more convolutional layers is often needed to improve model performance. However, deeper networks suffer from the vanishing gradient problem, where the gradient is diminished as it is backpropagated through multiple layers [80], [81]. Trainable parameters in the earlier layers may fail to converge to optimal values resulting in suboptimal model performance. Dense connectivity addresses this problem by introducing numerous concatenation connections

between convolutional layers which enable gradient information to directly flow into earlier layers [65].

In a dense block, the goal is to learn a total of $\boldsymbol{f}$ features from the input features. This is achieved by iterating through several steps, where $\boldsymbol{k}$ additional features are learned at each step. The key feature of dense connectivity is that earlier convolutional layers are connected to all subsequent layers by channel-wise concatenation [65], [72]. Each successive step learns additional features based on the original input provided and other features learned in previous layers. This removes the need to learn redundant features and promotes learning a more diverse set of features.

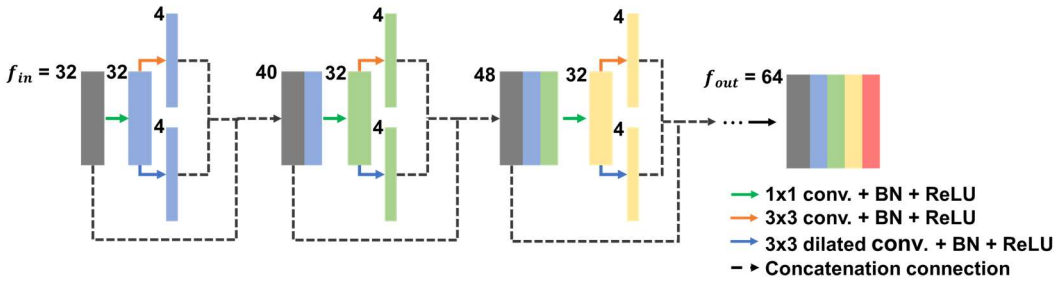

Fig. 24 Four layered dense dilation block with $k = 8$ and $f = 64$. In a dense dilation block, features learned from each convolutional layer are concatenated together with the input. Features are learned using both the standard and dilated convolution.

In this work, the dense block was modified to use both the standard and dilated convolutions (Fig. 24). At each step, $k/2$ features are learned with a standard convolution and the remaining features are learned using dilated convolutions with a dilation rate, $\boldsymbol{r}$. This combination was used to mitigate potential gridding artifacts that may arise from

using solely dilated convolutions. Furthermore, dilated convolutions, having a larger receptive field, can efficiently learn global context. Whereas, standard convolutions, having a denser receptive field, can efficiently learn local context.

Dense Dilation Blocks

The DD-Net is an enhanced version of our previous work, the Fully Dense U-Net (FD-UNet), which was shown to be superior to the standard UNet. The DD-Net follows an "encoder-decoder-refinement" structure [66]. The key innovation in the DD-Net is the unique use of dense dilation blocks to leverage the benefits of dense connectivity and dilated convolutions (Fig. 25). Max pooling layers were removed in the DD-Net because they often result in high frequency content that may cause gridding artifacts [109]. They were replaced by 2x2 convolutional layers with a stride of two which allow the CNN to learn a more useful transformation for spatial down sampling. A shallow "refinement" network comprised of a dense dilation block and two 3x3 convolutional layers was also added to the end of the original network. These additional layers allow the CNN to further refine the image and remove artifacts at the highest spatial resolution. In the original FD-UNet, only a 1x1 convolutional layer was applied at the end of the decoding stage to form the final image. Addition of a "refinement" stage has been shown to improve model performance [113].

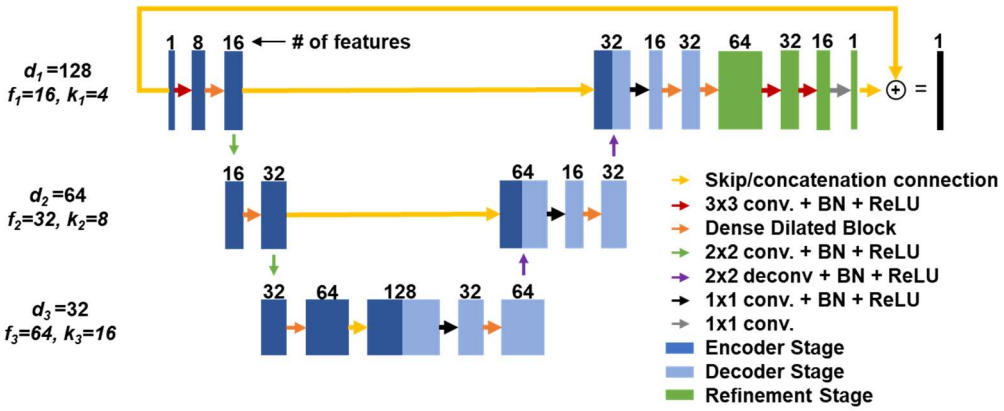

Fig. 25 Proposed DD-UNet architecture that incorporates dense connectivity and dilated convolutions throughout the UNet. In addition to the encoder and decoder structure of the standard UNet, several convolutional layers collectively termed the "refinement stage" were included following the decoder stage. Hyperparameters for the illustrated architecture are $k_1 = 8$ and $f_1 = 16$ for an input image of size 128x128x128 pixels.

Generating Training and Testing Data

Synthetic sphere phantoms were generated by placing 25-50 spheres with randomly selected center coordinates, radius (range 5 to 10 pixels), and magnitude (range 1 to 5) in a 128x128x128 pixels image. Resulting images were smoothed with a 5x5 moving average filter. This process was repeated to create a training dataset with 1000 images and a testing dataset with 500 images.

The "ELCAP Public Lung Image Database" is comprised of 50 whole-lung CT scans that were obtained within a single breath hold [101]. These scans were split into training (N=40) and testing groups (N=10) and were used to generate additional training and testing data via data augmentation. Each 3D scan had dimensions of 512x512x288 pixels. First, the lungs were segmented from the CT scan using active contours with the Chan-Vese algorithm [114]. Next, the Frangi vesselness filter was applied to suppress background noise and segment vessel-like structures in the lungs [86]. 3D vasculature

phantoms were then procedurally generated by randomly rotating the filtered 3D images along each axis and then sampling a 128x128x128 pixels image. This data augmentation process was repeated to create a training dataset with 1000 images and a testing dataset with 500 images.

Synthetic breast vasculature phantoms were created using an analytic approach to generate random but realistic anatomical structures within a predefined breast volume [115]. This method was originally developed for the "Simulated Virtual Imaging Clinical Trial for Regulatory Evaluation" project which sought to demonstrate *in silico* imaging trials and imaging computer simulation tools as a viable source of evidence for the regulatory evaluation of imaging devices. From this approach, 400 different breast phantoms with dimensions of 718x796x506 were generated. These phantoms were split into training (N=300) and testing groups (N=100) and were used to generate additional training and testing data via a similar data augmentation strategy as described earlier. A training dataset with 1000 images and a testing dataset with 500 images of breast vasculature were created.

The MATLAB toolbox k-WAVE was used to simulate photoacoustic data acquisition using an array of acoustic sensors arranged in a cylindrical geometry [83]. The sensor array is essentially a linear array with 128 elements along the z-axis that is repeated at equally spaced intervals along a half-circle in the x-y plane (Fig. 21a). In order to have experiments with varying levels of sparsity, three different sensor arrays with sampling at 10, 20, 30 angles in the x-y plane were used for simulations. Having fewer angles sampled results in more severe sparse spatial sampling artifacts.

Training and testing phantoms were normalized (values between 0 and 1) and treated as a photoacoustic source distribution on a computation grid of 128x128x128 pixels. The medium was assumed to be non-absorbing and heterogeneous, in which the background had a speed of sound of 1480 m/s and density of 1000 kg/m^3 while the vasculature had a speed of sound of 1570 m/s and density of 1060 kg/m^3. The time reversal method in the k-WAVE toolbox was used for reconstructing an initial image from the simulated photoacoustic time series data. In the scenario of *in vivo* imaging, the spatial distribution of the speed of sound and density is unknown. Thus, the reconstruction was completed assuming a homogeneous medium with a speed of sound of 1480 m/s and density of 1000 kg/m^3.

Evaluating Image Quality for Sparse Images

To evaluate image reconstruction quality, the multi-scale structural similarity index metric (MS-SSIM) was used to compare the reconstructed image to the ground truth image [116]. MS-SSIM is a composite metric that measures similarities between two images in terms of contrast, luminance, and structure at multiple spatial scales. Similarities are calculated based on a local neighborhood of pixels, and a global value is reported by averaging the neighborhood values. MS-SSIM is superior to other metrics such as the standard SSIM and peak-signal-to-noise-ratio (PSNR) for evaluating image quality in 3D images with sparse imaging targets that occupy a small fraction of the space in the medium. The main drawback of the SSIM and PSNR metrics is that image quality is only evaluated at a single spatial scale. Therefore, these metrics are heavily biased by

how well artifacts were removed from the background and only weakly associated with how well sparse structures were reconstructed.

Deep Learning Implementation

The CNNs are implemented in Python 3.7 with TensorFlow v2.1, an open source library for deep learning [87]. Training and evaluation of the network is performed on an NVIDIA V100 GPU. The CNNs were trained using the Adam optimizer to minimize the mean squared error loss with an initial learning rate of 1e-4 and a batch size of two images for 500 epochs. The same hyperparameters (i.e., f=16, k=4, and L=3) were used for both CNNs, and the DD-UNet had a dilation rate of two. Training each CNN required approximately one day to complete. A separate CNN was trained for each CNN architecture, PAT imaging system configuration, and dataset. The FD-UNet (120,000) and DD-UNet (150,000) had a similar number of parameters.

Section Three: Results

In silico experiments were performed using three different sparse imaging phantoms (i.e., spheres, lung vasculature, and breast vasculature) to evaluate the FD-UNet and DD-UNet for Post-DL image reconstruction. Given an initial time reversal reconstructed image, the CNNs were tasked with removing artifacts arising from sparse spatial sampling, limited-view detection, and an unknown heterogeneous medium. The MS-SSIM metric was used to evaluate image quality by comparing the reconstructed images to the ground truth image for N=500 testing image pairs in each dataset. Using

Otsu's method for automated thresholding and binarization, the imaging targets were estimated to on average occupy 3-5% of the space in the imaging phantoms.

Visual Comparison of CNN Images

In this initial experiment, the imaging system used a half-view cylindrical sensor array to sample the acoustic waves at 30 equally spaced angles. In general, it is difficult to visually identify differences between the FD-UNet and DD-UNet image reconstructions because the differences are subtle. Maximum intensity projections are convenient for visualizing 3D features in a 2D image, but only differences between the most prominent features can be seen in the projections. In the representative examples, both CNNs produce images that are of higher quality than the time reversal reconstructed images (Fig. 26). Spheres and vessels not visible in the time reversal reconstructions can be clearly seen in the CNN reconstructions. While both CNNs remove most artifacts and accurately reconstruct the larger and more prominent image features. The DD-UNet was observed to be better in reconstructing the smaller image features. These features are typically either missing or inaccurately reconstructed in the FD-UNet images. Furthermore, the FD-UNet occasionally mistakenly interpreted artifacts observed in the time reversal reconstruction as a true imaging target.

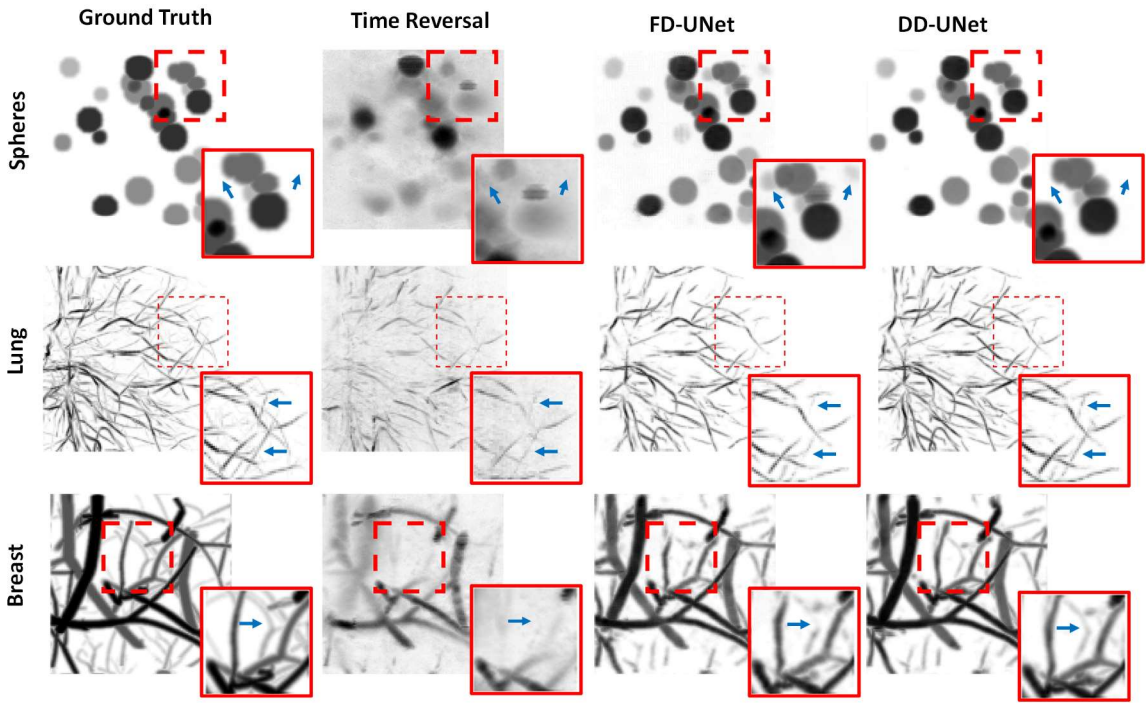

Fig. 26 Example ground truth and reconstructed images using the time reversal, FD-UNet, and the DD-UNet (dilation rate = 2) methods for three different imaging phantoms reconstructed with a sampling sparsity of 30 angles. The smaller image with a solid red border is an enlarged sub-image from the region designated by the dashed red line. The blue arrows highlight key differences between the reconstructed images (Top) Spheres phantom. The FD-UNet image incorrectly had spheres in the background that were not in the ground truth or DD-UNet images. (Middle) Lung vasculature phantom. The small vessels were more visible and clearer in the DD-UNet image than the FD-UNet image. (Bottom) Breast vasculature phantom. A small vessel that was not in the time reversal image was recovered in the CNN images but appeared to be sharper in the DD-UNet image.

Quantitative Comparison of CNN Performance

Images reconstructed with the CNNs had MS-SSIM scores ranging from 0.85 to 0.97 (spheres), 0.59 to 0.88 (lung), and 0.57 to 0.88 (breast). More complex imaging phantoms (e.g., containing more spheres or vessels) or those with imaging targets further away from the imaging sensor array typically resulted in reconstructions with lower scores. For all experiments performed, the DD-UNet consistently outperformed the FD-UNet when comparing the MS-SSIM for the same image reconstructed (Fig. 26). There

were only a few instances in which the FD-UNet reconstructed the test image with a higher MS-SSIM than DD-UNet. Interestingly, the degree of improvement in MS-SSIM by the DD-UNet appeared to have a stochastic nature. When examining the distribution of differences in MS-SSIM between the CNNs, the DD-UNet outperformed the FD-UNet with a mean and standard deviation of 0.033 ± 0.016 (spheres), 0.017 ± 0.009 (lung), and 0.027 ± 0.015 (breast) (Fig. 27).

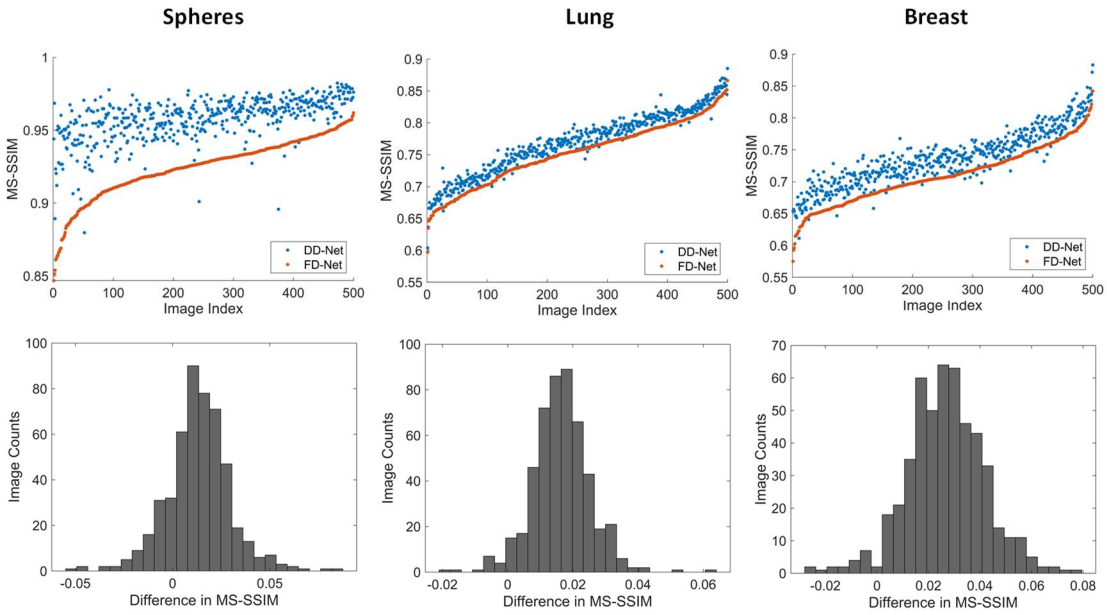

Fig. 27 (Top) Scatter plots for comparing the MS-SSIM of the FD-UNet and DD-UNet (dilation rate = 2) image reconstructions for each imaging phantom. For improved visualization, the image index was defined based on the sorted order of the MS-SSIM scores for the FD-UNet. (Bottom) Histogram showing the difference in MS-SSIM for the same test image between the DD-UNet and FD-UNet for each imaging phantom. A positive difference indicates that the DD-UNet reconstructed a higher quality image. Results shown are for a sparsity level of 30 angles sampled.

CNN Performance at Different Levels of Sparsity

By decreasing the number of angles sampled, the acoustic waves were more sparsely sampled. This resulted in increasingly severe streaking artifacts and thus in a more difficult problem for the CNNs to overcome. As expected, the average MS-SSIM scores decreased as the number of angles sampled decreased for all reconstruction methods (Table I). Both CNNs consistently improved the time reversal reconstruction for all imaging phantoms and levels of sparsity tested. The large difference in MS-SSIM between the CNNs and time reversal reconstructions can be mostly explained by the fact that both CNNs were highly proficient at removing background artifacts and properly reconstructing the larger image features (Fig. 24). The DD-UNet was shown to significantly outperform the FD-UNet for all imaging phantoms and levels of sparsity tested (Wilcoxon matched-pairs signed rank test, $p<0.01$).

Section Four: Discussion and Conclusion

In this work, we propose a modified CNN architecture termed DD-Net for 3D sparse and limited-view PAT image reconstruction that leverages the benefits of both dense connectivity and dilated convolutions through the unique use of dense dilation blocks. *In silico* experiments were performed with three different sparse phantoms (i.e., spheres, lung vasculature, and breast vasculature), and the DD-Net was demonstrated to be a superior CNN architecture compared to the FD-UNet. For all experiments performed, the DD-Net consistently reconstructed the image with a higher MS-SSIM by 0.01 to 0.03 depending on the phantom and level of sampling sparsity. Images

reconstructed by the DD-UNet and FD-UNet did not have many large and obvious visual differences. However, the DD-UNet was observed to be able to reconstruct the smaller structures and finer details more accurately. For example, small vessels in the breast and lung vasculature phantoms that were missing or inaccurately reconstructed in the FD-UNet image could be seen more clearly in the DD-Net image (Fig. 24). These improvements were likely due to the expanded receptive field enabling the CNN to use more context in the image to reconstruct these finer features and the addition of a shallow network termed the refinement stage to further correct artifacts at the highest image resolution.

Choice of dilation rate for the DD-UNet depends on the imaging targets to be reconstructed and size of the imaging volume. For example, larger volumes with predominantly bigger image features may benefit from a larger dilation rate since more global context is available to the CNN. However, increasing the dilation rate does not necessarily lead to improved performance since gridding artifacts can become more severe. Some image features might also be smaller than the zero-filled gaps in a large receptive field leading to a loss of local context, but this issue is mitigated to a degree in the DD-UNet by using a combination of standard and dilated convolutions.

A limitation in applying deep learning for 3D PAT image reconstruction is the limited GPU memory. This forces the use of shallower CNNs with fewer convolutional layers, which constrains the representational power or complexity of the CNN and results in suboptimal model performance. For example, the CNNs in this work for 3D PAT have about $\sim 10^5$ parameters due to memory limitations, while an equivalent CNN for 2D PAT

had ~10^6 or more parameters. Reconstructing larger volumes requires more memory and further limit the complexity of the CNN. Further work in developing more memory efficient CNN architecture or strategies for 3D PAT image reconstruction is needed to address this issue.

A key challenge in applying deep learning for *in vivo* PAT image reconstruction is the need for a large training dataset. Arbitrarily large synthetic training data can be generated using numerical phantoms and anatomical templates with data augmentation as demonstrated in this work. The synthetic data does need to properly capture the expected variations in artifacts observed in the experimental data. Therefore, the PAT simulation parameters need to be well-matched with the experimental conditions. Depending on the PAT imaging system, this can be a non-trivial task since it requires the system to be well-characterized.

The proposed Post-DL approach using the DD-Net can be used as a computationally efficient method for improving PAT image quality under limited-view and sparse sampling conditions. It can be applied to a wide variety of PAT imaging applications and allows for the development of more efficient data acquisition using fewer sensors without sacrificing image quality. This approach enables real-time PAT image rendering which would provide valuable feedback while imaging. The DD-Net can also be readily applied to image reconstruction problems in other imaging modalities (e.g., ultrasound and CT) and other biomedical imaging applications such as segmentation.

CHAPTER SIX
FOURIER NEURAL OPERATORS

Model-based learning has been demonstrated to outperform other deep learning frameworks. By incorporating an explicit model of photoacoustic wave propagation into the image reconstruction process, data consistency between the measured time-series data and the reconstructed image is greatly improved. However, this comes at the cost of increased computational cost and complexity since the forward and adjoint operators need to be repeatedly evaluated. A natural next step in the development of model-based learning is to improve its computational efficiency, which led to the idea of using neural networks as computational efficient approximations of the forward and adjoint operators. In other words, the neural network is used to solve the underlying partial differential equations that govern photoacoustic wave propagation. In this work, we employed Fourier Neural Operator (FNO) networks to approximate the forward operator and demonstrated as a proof-of-concept that is capable of accurately solving the underlying partial differential equations. At the time of completing this dissertation, the results of the FNO network are being considered for publication and are under review.

Section One: Introduction and Motivation

PAT simulation is a highly useful tool that provides quantitative and qualitative insights into these parameters affecting image quality [83]. It is commonly used prior to experimentation and imaging to optimize the system configuration. It also plays an integral role in image reconstruction and provides numerical phantom data for the

development of advanced algorithms such as iterative methods and deep learning methods [45], [61], [63], [107], [117], [118]. Simulating the PAT image acquisition is comprised of two components, the optical illumination and photoacoustic propagation. For this work, we are primarily focused on the photoacoustic component. The equation for photoacoustic wave propagation can be solved numerically using classical methods such as the time domain finite element method [119], [120]. However, these methods can become computationally expensive, especially for large three-dimensional (3D) simulations.

Recently, deep learning has been explored as a computationally efficient partial differential equation (PDE) solver [121], [122]. It has the potential to revolutionize scientific disciplines and research by providing fast PDE solvers that approximate or enhance conventional ones. Applications requiring repeated evaluations of the forward model can greatly benefit from having reduced computation times. Here, we provide a brief overview of three deep learning methods for solving PDEs – finite dimensional operators, neural finite element models, and Fourier neural operators (FNO).

Finite dimensional operators use a deep convolutional neural network (CNN) to solve the PDE on a finite Euclidean Space [98], [123]. By definition, this approach is mesh-dependent, and the CNN needs to be retrained for solving the PDE at different resolutions and discretization. Neural finite element models are mesh-independent and closely resembles traditional finite element methods [121], [124]. It replaces the set of local basis functions in the finite element models with a fully connected neural network. It requires prior knowledge of the underlying PDE and is designed to solve for one

specific instance of the PDE. The neural network needs to be retrained for new instances where the underlying PDE is parameterized with a different set of functional coefficients. FNO is a mesh-free approach that approximates the mapping between two infinite dimensional spaces from a finite collection of input-output paired observations [125], [126]. The neural operator is learned directly in the Fourier Space using a CNN. The same learned operator can be used without retraining to solve PDEs with different discretization and parameterization. Fourier Neural Operators have been demonstrated to achieve state-of-the-art results for a variety of PDEs (e.g., Burger's equation, Darcy Flow, and Navier-Stokes) and outperformed other existing deep learning methods [126].

To the best of our knowledge, this is the first paper that seeks to apply deep learning for solving the photoacoustic wave equation for simulating PAT. FNOs were chosen for this task given its flexibility in discretization and superior performance compared to other deep learning methods. Our contributions include adapting the FNO neural network and applying it as a fast PDE solver for simulating 2D photoacoustic wave propagation. Simulations from the FNO network and the widely used k-Wave toolbox for time domain acoustic wave propagation were compared in terms of accuracy and computation times. Further experiments were also conducted to evaluate the generalizability of the FNO network beyond the training data and the impact of key hyperparameters on network performance and complexity.

Section Two: Methods

Numerical approaches such as the finite-difference and finite-element methods are commonly used to solve PDEs by discretizing the space into a grid [127]. However, these methods are often slow for time domain modeling of broadband or high-frequency waves due to the need for a fine grid with small time-steps [83]. Computational efficiency can be improved using pseudo-spectral and k-space methods. The pseudo-spectral method fits a Fourier series to the data and reduces the number of grid points per wavelength required for an accurate solution [128]. The k-space method incorporates *a priori* information regarding the governing wave equation into the solution [129]. This allows for larger time steps and improves numerical stability in the case of acoustically heterogeneous mediums. The k-Wave toolbox, a widely used MATLAB tool for photoacoustic simulations, uses the pseudo-spectral k-space approach for solving time-domain photoacoustic wave simulations [130].

Fourier Neural Operator Network

The FNO network was adapted for solving the 2D photoacoustic wave equation [126]. In our version, the FNO network does not apply Gaussian normalization to either the input or output of the training example. The network begins by mapping the input into a higher dimensional representation using a fully connected layer (Fig. 28). The transformed features are then iteratively updated by passing them through four successive Fourier layers. Finally, the updated features are projected to the desired dimensions using a fully connected layer. Through a combination of linear, Fourier, and non-linear

transformations, the Fourier neural operators can approximate complex operators in PDEs that are highly non-linear with high frequency modes.

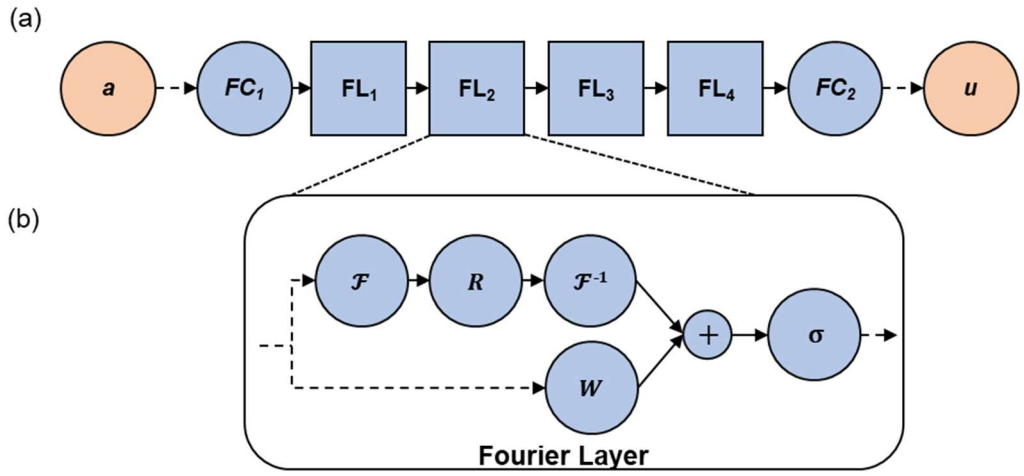

Fig. 28 (a) Neural network architecture for the FNO network. The input *a* is mapped to a higher dimensional space using a fully connected layer (FC_1). The transformed feature is passed through four Fourier Layers (FL). Finally, a fully connected layer (FC_2) is used to obtain the final output *u* with the desired dimensions. (b) Architecture of a Fourier layer. The input goes through two paths in the Fourier layer. In the top path, the input undergoes a Fourier Transform $\mathcal{F}$, linear transform *R*, and inverse Fourier Transform $\mathcal{F}^{-1}$. In the bottom path, the input undergoes a linear transform *W*. Outputs from each path are summed together and undergo ReLU activation σ.

The photoacoustic wave equation can be solved with the FNO network using either a 2D or 3D implementation. In 2D, the FNO network performs 2D convolutions in space and finds a solution for some fixed interval length Δt. The solution is then recurrently propagated in time and used to solve for the next interval length. In 3D, the FNO network performs 3D convolutions in space-time and can directly output the full time series solution with any time discretization. While both implementations were demonstrated to have similar performance, the 3D FNO network was used in this work because it was found to be more expressive and easier to train [126]. The FNO network

was implemented in Python v3.8 using the deep learning library PyTorch v1.7.1. The Adam optimizer with a mean squared error loss function was used to train the FNO network for 2,000 epochs over approximately two days on a NVIDIA Tesla K80 graphics processing unit (GPU).

Channels and modes are the two main hyperparameters that impact the accuracy of the FNO network. The channels parameter defines the width of the FNO network meaning the number of features learned in each layer. The modes parameter defines the number of lower Fourier modes retained when truncating the Fourier series. The allowable maximum number of modes is related to the size of the simulation computational grid. In this work, the FNO network is assumed to have 64 modes and 5 channels unless otherwise specified.

Photoacoustic Data for Training and Testing

The MATLAB toolbox k-Wave was used for photoacoustic wave simulation and to generate data for training and testing the FNO network [83]. The simulation medium was defined as a 64x64 computational grid, non-absorbing, and homogenous with a speed of sound of 1480 m/s and density of 1000 kg/m^3. Simulations were performed with a timestep of 20 ns for T=151 steps. The initial photoacoustic pressure was initialized using anatomically realistic breast vasculature phantoms that were numerically generated [115]. The training dataset (N=500) and testing dataset (N=100) were comprised of images representing the initial photoacoustic pressure (the input to the FNO network), and the corresponding simulation of the photoacoustic wave propagation (output of the FNO

network). The FNO output was compared against the photoacoustic simulation performed using k-Wave, which served as the ground truth. While the training and testing data shared similar features, each example was unique. Simulations for the Shepp-Logan, synthetic vasculature, tumor, and Mason-M phantoms were also generated to evaluate the generalizability of the FNO network [32], [83].

Section Three: Results

Comparison of FNO Network and k-Wave Simulations

When tested on breast-vascular images like those used for training, the photoacoustic-wave simulations produced by the FNO network and k-Wave were remarkably similar and almost visually identical (Fig. 29). This demonstrated that the FNO network can model both broadband and high-frequency waves required for photoacoustic simulations. The FNO network and k-Wave simulations were quantitatively compared using the mean squared error (MSE). For the testing dataset, the MSE of the FNO network was 3.1e-5 which indicates that the FNO network was able to accurately simulate photoacoustic wave propagation.

The time required to solve the photoacoustic wave equation largely depends on the discretization of the computational grid. In k-Wave, a simulation using a 64x64 grid required ~1.17 seconds to complete on a GPU. For a comparable simulation, the FNO network only required ~0.029 seconds to complete on a GPU which is approximately a 40x reduction in computation time.

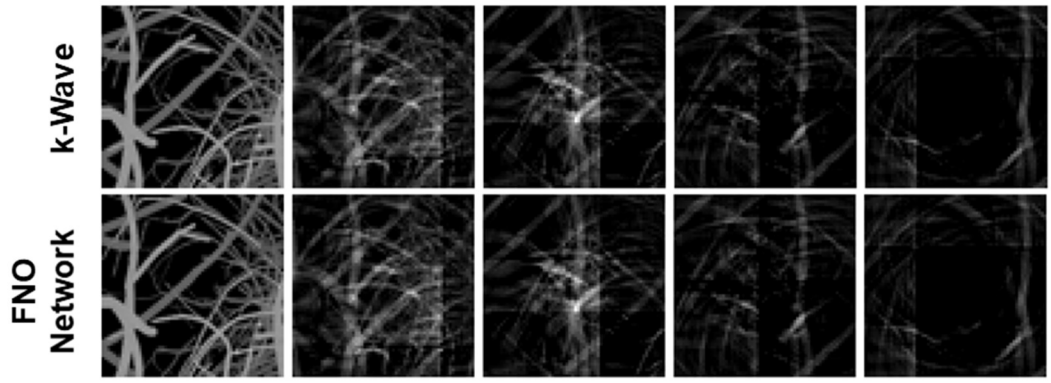

Fig. 29 Visual comparison of the ground truth (Top Row) using k-Wave and the FNO network (Bottom Row) simulated photoacoustic wave propagation for an example vasculature image in a homogeneous medium at T = 1, 20, 40, 60, and 80 timesteps. The MSE for this example was 2.5e-5.

PAT Images Reconstructed from Simulations

For further validation, an *in-silico* experiment of PAT imaging with a 64-sensor linear array was conducted using the k-Wave and the FNO network simulations. Other sensor arrays and geometries can be used, but the linear array was chosen since it is widely available and used in laboratories. The sensor data for image reconstruction was created by sampling the photoacoustic pressures along the top row of the computational grid in each simulation. Images were then reconstructed from the time-series sensor data using the time reversal method in k-Wave [83]. The reconstructed images were highly similar with only minor differences (Fig. 30). The vasculature structures and limited-view artifact patterns seen in the FNO network image clearly matched those in the reconstructed image obtained using the k-Wave simulation data. The reconstructed images were quantitatively compared using MSE and the structural similarity index metric (SSIM), a metric ranging from 0 to 1 that measures the similarity between two images based on factors relevant to human visual perception (e.g., structure, contrast, and

luminance) [88]. For the testing dataset (N=100), the FNO network images had a MSE of 3.1e-5 and SSIM of 0.99. This demonstrated that the time-series sensor data produced using the FNO network and k-Wave simulations were effectively identical.

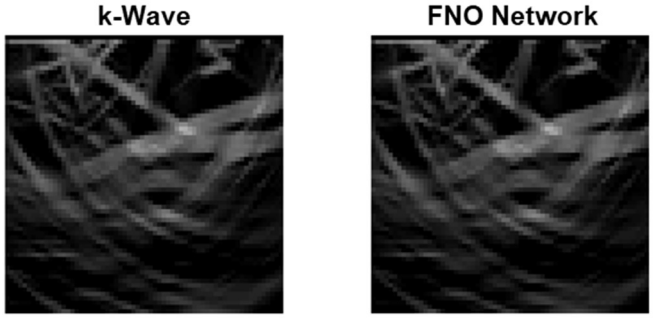

Fig. 30 Images reconstructed using sampled sensor data from the k-Wave and FNO network photoacoustic simulations. The images were normalized to have intensities between 0 and 1. For this example, the MSE and SSIM were respectively 6.1e-5 and 0.99.

FNO Network Generalizability

The FNO network was used to simulate photoacoustic wave propagation from Shepp-Logan, synthetic vasculature, tumor, and Mason-M phantoms. These phantoms contain many features not observed in the training dataset (breast vasculature). The FNO network and k-Wave simulations were visually similar for each phantom tested (Fig. 31). The MSE of the FNO network simulations were 2.1e-4 (Shepp-Logan), 3.7e-4 (synthetic vasculature), 1.9e-4 (tumor), and 6.4e-4 (Mason-M). These results provide evidence that the FNO network was generalizable to initial photoacoustic sources not in the training data.

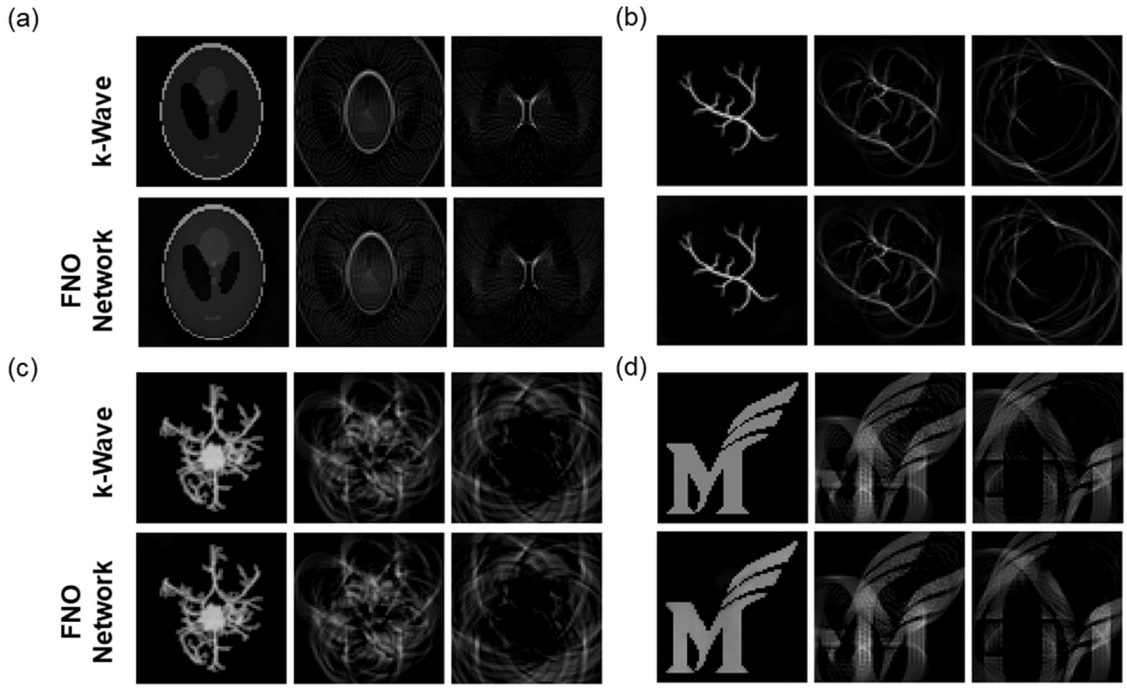

Fig. 31 Comparison between FNO Network and k-Wave simulations for initial pressure sources using the (a) Shepp-Logan, (b) synthetic vasculature, (c) tumor, and (d) Mason-M phantoms at T=1,10, and 20 timesteps.

Hyperparameter Optimization

A study was conducted to investigate the impact of hyperparameter selection on the FNO network's accuracy. The number of modes had the largest impact since it is directly related to the truncation error in a Fourier layer. Networks with a lower mode produced simulations with a blurred appearance due to the loss of high frequency information (Fig. 32). Increasing the number of channels generally improved the FNO network's performance but also required more GPU memory (Table 8). There was no benefit in having an FNO network with more than five channels. Interestingly, the computation time to complete a simulation was approximately the same for FNO networks with a lower number of modes or channels. There was a moderate increase in

computation time for the larger FNO networks with 64 modes and higher number of channels.

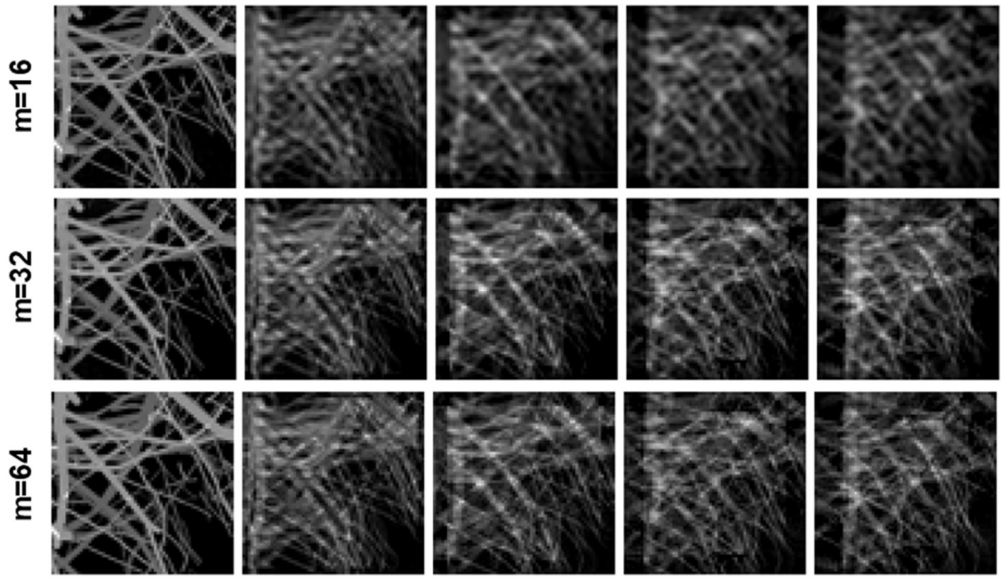

Fig. 32 Visual comparison of photoacoustic wave simulations at T = 1, 5, 10, 15, and 20 timesteps. The FNO networks were parametrized with channels=5 and modes=16, 32, and 64.

Table 8 Comparison of FNO networks for different hyperparameters

Modes	Channels	MSE	Time (s)	GPU Memory (GB)
16	5	1.0e-3	0.022	1.3
32	5	1.1e-4	0.019	1.7
64	5	3.1e-5	0.022	4.8
64	2	2.7e-4	0.022	1.8
64	3	1.3e-4	0.021	2.5
64	4	4.4e-5	0.021	3.5
64	5	3.1e-5	0.022	4.8
64	6	3.5e-5	0.023	6.3
64	7	2.8e-5	0.026	8.2
64	8	3.6e-5	0.028	10.2

Section Five: Discussion and Conclusion

Solving the 2D photoacoustic wave equation with traditional methods typically require a fine discretization of the computational grid and can be time-consuming to complete. Deep learning methods directly learn from data to solve PDEs and can be orders of magnitude faster with a minimal loss in accuracy. In this work, we applied the FNO network as a fast PDE solver for the 2D photoacoustic wave equation in a homogeneous medium. The FNO network and k-Wave solutions were qualitatively and quantitatively comparable. PAT images reconstructed from the FNO network and k-Wave simulations were effectively identical. This demonstrates that the sampled sensor data contained essentially the same information, and errors in the FNO network simulations did not impact the quality of images reconstructed. The FNO network was about 40x faster than k-Wave in completing a simulation with a 64x64 computational grid. Applications requiring repeated evaluations of the photoacoustic wave equation such as iterative image reconstruction can be accelerated using the FNO network.

Model generalizability is a highly desirable property because it removes the need to retrain the model when it is used on examples not observed in the training data. This is important since the goal of the FNO network is to be like traditional methods as a general PDE solver for any arbitrary initial pressure source. The FNO network's generalizability was evaluated by having it perform photoacoustic simulations for four phantoms not in the training data (e.g., Shepp-Logan, synthetic vasculature, tumor, and Mason-M). The FNO network and k-Wave simulations were highly similar indicating that a trained FNO network can be used for simulations with any arbitrary initial pressure source.

Furthermore, this provided evidence that the FNO network was learning the operator for photoacoustic wave propagation and not mainly specific solutions related to the training data.

Hyperparameter optimization is important to achieve the required level of accuracy and to minimize the memory required for training and inferring. The FNO network is parameterized by the number of modes and channels and increasing either parameter typically improves model performance. In general, a higher number of modes is preferred but can be reduced if only a lower-resolution approximation of the solution is needed. Hyperparameter optimization is likely more important for simulations with large computational grids when limited GPU memory can be a problem. Alternative network architectures that are more memory efficient such as the recurrent 2D FNO network can be explored [126].

In this work, the FNO network was trained for solving the 2D acoustic wave equation in a homogeneous medium. Simulations with homogeneous mediums are widely used in many applications such as image reconstruction where the spatial distribution of heterogeneities is often unknown. Nevertheless, the FNO network can be used for simulations with heterogeneous mediums. The spatial distribution of heterogeneous medium properties can be provided as an input to the FNO network. By providing training examples of simulations with varying heterogeneous mediums, the FNO network likely can learn to solve the 2D wave equation and account for effects due to the heterogeneous medium.

A practical limitation in data driven PDE solvers such as the FNO network is the need for high quality training data. Traditional solvers are often used to create arbitrarily large datasets to train the network. Depending on the size of the computational grid, this can be computationally formidable such as the case of 3D photoacoustic simulations. To create a large dataset in these scenarios, a high-performance computing environment would be needed to generate the training data in a reasonable timeframe.

CHAPTER SEVEN
FUTURE WORK AND DISCUSSION

In recent years, there has been a diverse body of work in literature demonstrating a wide variety of approaches in using deep learning for photoacoustic tomography image reconstruction problems. Fully learned reconstruction methods implicitly learn from data the physics of acoustic wave propagation to reconstruct an image from the measured time-series data. These methods have a low latency and are ideal for real-time imaging applications because it does not require a numerical model to solve the underlying physics. Other approaches like the learned iterative reconstruction uses an explicit physical model in combination with a neural network to provide state-of-the-art performance. Incorporating the physical model into the reconstruction process is highly useful in obtaining a more accurate and stable reconstruction but at the cost of increased computational complexity. There remains an open question of how and where to best use the physical model in relation with the neural network to obtain a fast, accurate, and robust reconstruction method. As demonstrated by the FNO networks, the physical model can be approximated using a neural network that is accurate and much faster than numerical methods. However, more work is needed to explore how well the FNO generalizes to heterogeneous mediums and 3D PAT simulations.

Data mismatch is an important challenge for data-driven methods like deep learning. Trained networks often do not perform well on data it has not previously observed in the training process. In many biomedical applications, it is challenging and often impossible to create a training dataset that contains examples for every possible

normal and abnormal image feature. Therefore, it is questionable if a relatively rare occurrence such as a small tumor in the image is real or an artifact of the neural network. Addressing this challenge is critical for the transition and building of trust in deep learning for clinical applications. Furthermore, many deep learning models rely on simulation tools such as k-Wave to generate large training datasets. Models trained on simulated data does not necessarily work well with in vivo and experimental data. It is difficult to ensure that the distribution of image features in the simulated data matches those of the in vivo data. While in vivo data can be acquired to supplement the training data, this is not a practical solution since the acquired data is instrumentation specific.

Deep learning has been successfully applied for many medical applications especially in the field of radiology [131]. However, there is still a critical question of how to integrate deep learning with PAT into clinical workflows. This likely depends on the specific target clinical application, but in general, it should be intuitive, easy-to-use, and provides value to the patient and clinicians [132]. While there is a multitude of promising clinical applications, a unique advantage of PAT over other conventional imaging modalities is its ability to measure blood perfusion and oxygenation [21]. This requires both the acoustic and optical inversions to be solved. Majority of the work in literature related to deep learning and PAT is focused on the acoustic inversion, but there has been some promising initial work for solving the optical inversion [133]. To fully realize the potential of deep learning and PAT for clinical applications, additional work solving the optical inversion is needed.

REFERENCES

[1] L. V. Wang, "Multiscale photoacoustic microscopy and computed tomography," *Nature Photon*, vol. 3, no. 9, pp. 503–509, Sep. 2009, doi: 10.1038/nphoton.2009.157.

[2] J. W. Goodman, "Some fundamental properties of speckle*," *J. Opt. Soc. Am., JOSA*, vol. 66, no. 11, pp. 1145–1150, Nov. 1976, doi: 10.1364/JOSA.66.001145.

[3] Z. Guo, L. Li, and L. V. Wang, "On the speckle-free nature of photoacoustic tomography," *Med Phys*, vol. 36, no. 9, pp. 4084–4088, Sep. 2009, doi: 10.1118/1.3187231.

[4] J. Xia and L. V. Wang, "Small-animal whole-body photoacoustic tomography: a review," *IEEE Trans Biomed Eng*, vol. 61, no. 5, pp. 1380–1389, May 2014, doi: 10.1109/TBME.2013.2283507.

[5] N. Nyayapathi and J. Xia, "Photoacoustic imaging of breast cancer: a mini review of system design and image features," *J Biomed Opt*, vol. 24, no. 12, Dec. 2019, doi: 10.1117/1.JBO.24.12.121911.

[6] B. L. Bungart *et al.*, "Photoacoustic tomography of intact human prostates and vascular texture analysis identify prostate cancer biopsy targets," *Photoacoustics*, vol. 11, pp. 46–55, Aug. 2018, doi: 10.1016/j.pacs.2018.07.006.

[7] C. Moore and J. V. Jokerst, "Strategies for Image-Guided Therapy, Surgery, and Drug Delivery Using Photoacoustic Imaging," *Theranostics*, vol. 9, no. 6, pp. 1550–1571, Feb. 2019, doi: 10.7150/thno.32362.

[8] M. Li, Y. Tang, and J. Yao, "Photoacoustic tomography of blood oxygenation: A mini review," *Photoacoustics*, vol. 10, pp. 65–73, Jun. 2018, doi: 10.1016/j.pacs.2018.05.001.

[9] L. V. Wang, "Prospects of photoacoustic tomography," *Med Phys*, vol. 35, no. 12, pp. 5758–5767, Dec. 2008, doi: 10.1118/1.3013698.

[10] P. Beard, "Biomedical photoacoustic imaging," *Interface Focus*, vol. 1, no. 4, pp. 602–631, Aug. 2011, doi: 10.1098/rsfs.2011.0028.

[11] J. Xia, J. Yao, and L. V. Wang, "Photoacoustic tomography: principles and advances," *Electromagn Waves (Camb)*, vol. 147, pp. 1–22, 2014.

[12] W. Li and X. Chen, "Gold nanoparticles for photoacoustic imaging," *Nanomedicine (Lond)*, vol. 10, no. 2, pp. 299–320, Jan. 2015, doi: 10.2217/nnm.14.169.

[13] D. Wu, L. Huang, M. S. Jiang, and H. Jiang, "Contrast Agents for Photoacoustic and Thermoacoustic Imaging: A Review," *Int J Mol Sci*, vol. 15, no. 12, pp. 23616–23639, Dec. 2014, doi: 10.3390/ijms151223616.

[14] M. Xu and L. V. Wang, "Universal back-projection algorithm for photoacoustic computed tomography," Apr. 2005, vol. 5697, pp. 251–255. doi: 10.1117/12.589146.

[15] B. E. Treeby, E. Z. Zhang, and B. T. Cox, "Photoacoustic tomography in absorbing acoustic media using time reversal," *Inverse Problems*, vol. 26, no. 11, p. 115003, 2010, doi: 10.1088/0266-5611/26/11/115003.

[16] L. V. Wang and S. Hu, “Photoacoustic Tomography: In Vivo Imaging from Organelles to Organs,” *Science*, vol. 335, no. 6075, pp. 1458–1462, Mar. 2012, doi: 10.1126/science.1216210.
[17] S. Hu and L. V. Wang, “Optical-resolution photoacoustic microscopy: auscultation of biological systems at the cellular level,” *Biophys J*, vol. 105, no. 4, pp. 841–847, Aug. 2013, doi: 10.1016/j.bpj.2013.07.017.
[18] S. Jeon, J. Kim, D. Lee, J. W. Baik, and C. Kim, “Review on practical photoacoustic microscopy,” *Photoacoustics*, vol. 15, p. 100141, Sep. 2019, doi: 10.1016/j.pacs.2019.100141.
[19] S. Park, C. Lee, J. Kim, and C. Kim, “Acoustic resolution photoacoustic microscopy,” *Biomed. Eng. Lett.*, vol. 4, no. 3, pp. 213–222, Sep. 2014, doi: 10.1007/s13534-014-0153-z.
[20] S. Hu, K. Maslov, and L. V. Wang, “Second-generation optical-resolution photoacoustic microscopy with improved sensitivity and speed,” *Opt Lett*, vol. 36, no. 7, pp. 1134–1136, Apr. 2011.
[21] A. B. E. Attia *et al.*, “A review of clinical photoacoustic imaging: Current and future trends,” *Photoacoustics*, vol. 16, p. 100144, Dec. 2019, doi: 10.1016/j.pacs.2019.100144.
[22] M. Xu and L. V. Wang, “Universal back-projection algorithm for photoacoustic computed tomography,” *Physical Review E*, vol. 71, no. 1, Jan. 2005, doi: 10.1103/PhysRevE.71.016706.
[23] S. Li, B. Montcel, W. Liu, and D. Vray, “Analytical model of optical fluence inside multiple cylindrical inhomogeneities embedded in an otherwise homogeneous turbid medium for quantitative photoacoustic imaging,” *Opt Express*, vol. 22, no. 17, pp. 20500–20514, Aug. 2014, doi: 10.1364/OE.22.020500.
[24] Y. Hristova, P. Kuchment, and L. Nguyen, “Reconstruction and time reversal in thermoacoustic tomography in acoustically homogeneous and inhomogeneous media,” *Inverse Problems*, vol. 24, no. 5, p. 055006, 2008, doi: 10.1088/0266-5611/24/5/055006.
[25] B. T. Cox and B. E. Treeby, “Artifact Trapping During Time Reversal Photoacoustic Imaging for Acoustically Heterogeneous Media,” *IEEE Transactions on Medical Imaging*, vol. 29, no. 2, pp. 387–396, Feb. 2010, doi: 10.1109/TMI.2009.2032358.
[26] Y. Xu, L. V. Wang, G. Ambartsoumian, and P. Kuchment, “Reconstructions in limited-view thermoacoustic tomography,” *Medical Physics*, vol. 31, no. 4, pp. 724–733, Apr. 2004, doi: 10.1118/1.1644531.
[27] A. Hauptmann and B. T. Cox, “Deep learning in photoacoustic tomography: current approaches and future directions,” *JBO*, vol. 25, no. 11, p. 112903, Oct. 2020, doi: 10.1117/1.JBO.25.11.112903.
[28] B. T. Cox, J. G. Laufer, P. C. Beard, and S. R. Arridge, “Quantitative spectroscopic photoacoustic imaging: a review,” *JBO*, vol. 17, no. 6, p. 061202, Jun. 2012, doi: 10.1117/1.JBO.17.6.061202.

[29] M. J. Willemink and P. B. Noël, "The evolution of image reconstruction for CT—from filtered back projection to artificial intelligence," *Eur Radiol*, vol. 29, no. 5, pp. 2185–2195, 2019, doi: 10.1007/s00330-018-5810-7.
[30] H. T. H. Piaggio, "The Mathematical Theory of Huygens' Principle," *Nature*, vol. 145, no. 3675, pp. 531–532, Apr. 1940, doi: 10.1038/145531a0.
[31] C. Huang, K. Wang, L. Nie, L. V. Wang, and M. A. Anastasio, "Full-Wave Iterative Image Reconstruction in Photoacoustic Tomography with Acoustically Inhomogeneous Media," *arXiv:1303.5680 [physics]*, Mar. 2013, Accessed: Aug. 30, 2018. [Online]. Available: http://arxiv.org/abs/1303.5680
[32] S. Arridge *et al.*, "Accelerated high-resolution photoacoustic tomography via compressed sensing," *Phys. Med. Biol.*, vol. 61, no. 24, p. 8908, 2016, doi: 10.1088/1361-6560/61/24/8908.
[33] A. Beck and M. Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, Jan. 2009, doi: 10.1137/080716542.
[34] A. Hauptmann *et al.*, "Model based learning for accelerated, limited-view 3D photoacoustic tomography," *arXiv:1708.09832 [cs, math]*, Aug. 2017, Accessed: Jul. 19, 2018. [Online]. Available: http://arxiv.org/abs/1708.09832
[35] J. Gu *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, May 2018, doi: 10.1016/j.patcog.2017.10.013.
[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
[37] G. Wang, J. C. Ye, K. Mueller, and J. A. Fessler, "Image Reconstruction is a New Frontier of Machine Learning," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1289–1296, Jun. 2018, doi: 10.1109/TMI.2018.2833635.
[38] J. Ye, Y. Han, and E. Cha, "Deep Convolutional Framelets: A General Deep Learning Framework for Inverse Problems," *SIAM J. Imaging Sci.*, pp. 991–1048, Jan. 2018, doi: 10.1137/17M1141771.
[39] E. Haber and L. Ruthotto, "Stable Architectures for Deep Neural Networks," *Inverse Problems*, vol. 34, no. 1, p. 014004, Jan. 2018, doi: 10.1088/1361-6420/aa9a90.
[40] L. Ruthotto and E. Haber, "Deep Neural Networks Motivated by Partial Differential Equations," *Journal of Mathematical Imaging and Vision*, vol. 62, Apr. 2018, doi: 10.1007/s10851-019-00903-1.
[41] K. Hammernik *et al.*, "Learning a variational network for reconstruction of accelerated MRI data," *Magnetic Resonance in Medicine*, vol. 79, no. 6, pp. 3055–3071, 2018, doi: 10.1002/mrm.26977.
[42] J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, and D. Rueckert, "A Deep Cascade of Convolutional Neural Networks for Dynamic MR Image Reconstruction," *IEEE Transactions on Medical Imaging*, vol. 37, no. 2, pp. 491–503, Feb. 2018, doi: 10.1109/TMI.2017.2760978.

[43] E. Kang, J. Min, and J. C. Ye, “A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction,” *Medical Physics*, vol. 44, no. 10, pp. e360–e375, Oct. 2017, doi: 10.1002/mp.12344.
[44] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, “Deep Convolutional Neural Network for Inverse Problems in Imaging,” *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4509–4522, Sep. 2017, doi: 10.1109/TIP.2017.2713099.
[45] A. Hauptmann *et al.*, “Model-Based Learning for Accelerated, Limited-View 3-D Photoacoustic Tomography,” *IEEE Transactions on Medical Imaging*, vol. 37, pp. 1382–1393, 2018, doi: 10.1109/TMI.2018.2820382.
[46] S. Antholzer, M. Haltmeier, R. Nuster, and J. Schwab, “Photoacoustic image reconstruction via deep learning,” in *Photons Plus Ultrasound: Imaging and Sensing 2018*, Feb. 2018, vol. 10494, p. 104944U. doi: 10.1117/12.2290676.
[47] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
[48] J. Schmidhuber, “Deep learning in neural networks: An overview,” *Neural Networks*, vol. 61, pp. 85–117, Jan. 2015, doi: 10.1016/j.neunet.2014.09.003.
[49] F. Emmert-Streib, Z. Yang, H. Feng, S. Tripathi, and M. Dehmer, “An Introductory Review of Deep Learning for Prediction Models With Big Data,” *Front. Artif. Intell.*, vol. 0, 2020, doi: 10.3389/frai.2020.00004.
[50] L. Alzubaidi *et al.*, “Review of deep learning: concepts, CNN architectures, challenges, applications, future directions,” *Journal of Big Data*, vol. 8, no. 1, p. 53, Mar. 2021, doi: 10.1186/s40537-021-00444-8.
[51] S. Indolia, A. K. Goswami, S. P. Mishra, and P. Asopa, “Conceptual Understanding of Convolutional Neural Network- A Deep Learning Approach,” *Procedia Computer Science*, vol. 132, pp. 679–688, Jan. 2018, doi: 10.1016/j.procs.2018.05.069.
[52] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, “Convolutional neural networks: an overview and application in radiology,” *Insights Imaging*, vol. 9, no. 4, Art. no. 4, Aug. 2018, doi: 10.1007/s13244-018-0639-9.
[53] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: 10.1038/323533a0.
[54] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” Dec. 2014, Accessed: Jul. 29, 2021. [Online]. Available: https://arxiv.org/abs/1412.6980v9
[55] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, “A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science,” in *Supervised and Unsupervised Learning for Data Science*, M. W. Berry, A. Mohamed, and B. W. Yap, Eds. Cham: Springer International Publishing, 2020, pp. 3–21. doi: 10.1007/978-3-030-22475-2_1.
[56] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks,” Mar. 2017, Accessed: Jul. 29, 2021. [Online]. Available: https://arxiv.org/abs/1703.10593v7

[57] H. Deng, H. Qiao, Q. Dai, and C. Ma, “Deep learning in photoacoustic imaging: a review,” *JBO*, vol. 26, no. 4, p. 040901, Apr. 2021, doi: 10.1117/1.JBO.26.4.040901.
[58] J. Gröhl, M. Schellenberg, K. Dreher, and L. Maier-Hein, “Deep learning for biomedical photoacoustic imaging: A review,” *arXiv:2011.02744 [physics]*, Nov. 2020, Accessed: Nov. 18, 2020. [Online]. Available: http://arxiv.org/abs/2011.02744
[59] Waibel, Dominik, Grohl, Janek, Isensee, Fabian, Kirchner, Thomas, Maier-Hein, Klaus, and Maier-Hein Lena, “Reconstruction of initial pressure from limited view photoacoustic images using deep learning,” in *Photons Plus Ultrasound: Imaging and Sensing 2018*, San Francisco, CA, vol. 10494. Accessed: Oct. 28, 2018. [Online]. Available: https://www-spiedigitallibrary-org.mutex.gmu.edu/conference-proceedings-of-spie/10494/104942S/Reconstruction-of-initial-pressure-from-limited-view-photoacoustic-images-using/10.1117/12.2288353.full?SSO=1
[60] E. M. A. Anas, H. K. Zhang, C. Audigier, and E. M. Boctor, “Robust Photoacoustic Beamforming Using Dense Convolutional Neural Networks,” in *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation*, Cham, 2018, pp. 3–11. doi: 10.1007/978-3-030-01045-4_1.
[61] S. Antholzer, M. Haltmeier, and J. Schwab, “Deep learning for photoacoustic tomography from sparse data,” *Inverse Problems in Science and Engineering*, vol. 27, no. 7, pp. 987–1005, Jul. 2019, doi: 10.1080/17415977.2018.1518444.
[62] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” *arXiv:1505.04597 [cs]*, May 2015, Accessed: Feb. 08, 2018. [Online]. Available: http://arxiv.org/abs/1505.04597
[63] S. Guan, A. Khan, S. Sikdar, and P. Chitnis, “Fully Dense UNet for 2D Sparse Photoacoustic Tomography Artifact Removal,” *IEEE J Biomed Health Inform*, Apr. 2019, doi: 10.1109/JBHI.2019.2912935.
[64] S. Antholzer, M. Haltmeier, and J. Schwab, “Deep Learning for Photoacoustic Tomography from Sparse Data,” *arXiv:1704.04587 [cs]*, Apr. 2017, Accessed: Sep. 24, 2017. [Online]. Available: http://arxiv.org/abs/1704.04587
[65] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” *arXiv:1608.06993 [cs]*, Aug. 2016, Accessed: Apr. 11, 2018. [Online]. Available: http://arxiv.org/abs/1608.06993
[66] 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*, 2015, pp. 234–241.
[67] M. Haltmeier, “Sampling Conditions for the Circular Radon Transform,” *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2910–2919, Jun. 2016, doi: 10.1109/TIP.2016.2551364.
[68] A. Rosenthal, V. Ntziachristos, and D. Razansky, “Acoustic Inversion in Optoacoustic Tomography: A Review,” *Curr Med Imaging Rev*, vol. 9, no. 4, pp. 318–336, Nov. 2013, doi: 10.2174/15734056113096660006.
[69] J. Frikel and M. Haltmeier, “Efficient regularization with wavelet sparsity constraints in PAT,” *arXiv:1703.08240 [math]*, Mar. 2017, Accessed: Mar. 08, 2018. [Online]. Available: http://arxiv.org/abs/1703.08240

[70] K. Wang, R. Su, A. A. Oraevsky, and M. A. Anastasio, “Investigation of iterative image reconstruction in three-dimensional optoacoustic tomography,” *Phys. Med. Biol.*, vol. 57, no. 17, p. 5399, 2012, doi: 10.1088/0031-9155/57/17/5399.

[71] Y. Han, J. Yoo, and J. C. Ye, “Deep Residual Learning for Compressed Sensing CT Reconstruction via Persistent Homology Analysis,” Nov. 2016.

[72] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P. A. Heng, “H-DenseUNet: Hybrid Densely Connected UNet for Liver and Liver Tumor Segmentation from CT Volumes,” *arXiv:1709.07330 [cs]*, Sep. 2017, Accessed: Apr. 11, 2018. [Online]. Available: http://arxiv.org/abs/1709.07330

[73] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation,” *arXiv:1606.06650 [cs]*, Jun. 2016, Accessed: Sep. 24, 2017. [Online]. Available: http://arxiv.org/abs/1606.06650

[74] J. Schwab, S. Antholzer, R. Nuster, and M. Haltmeier, “DALnet: High-resolution photoacoustic projection imaging using deep learning,” *arXiv:1801.06693 [physics]*, Jan. 2018, Accessed: Aug. 22, 2018. [Online]. Available: http://arxiv.org/abs/1801.06693

[75] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” *arXiv:1502.03167 [cs]*, Feb. 2015, Accessed: Mar. 31, 2018. [Online]. Available: http://arxiv.org/abs/1502.03167

[76] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, “How Does Batch Normalization Help Optimization? (No, It Is Not About Internal Covariate Shift),” *arXiv:1805.11604 [cs, stat]*, May 2018, Accessed: Aug. 27, 2018. [Online]. Available: http://arxiv.org/abs/1805.11604

[77] Z. Zhang, X. Liang, X. Dong, Y. Xie, and G. Cao, “A Sparse-View CT Reconstruction Method Based on Combination of DenseNet and Deconvolution,” *IEEE Transactions on Medical Imaging*, vol. 37, pp. 1–1, Apr. 2018, doi: 10.1109/TMI.2018.2823338.

[78] J. Long, E. Shelhamer, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” p. 10.

[79] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. Accessed: Mar. 08, 2018. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[80] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *arXiv:1512.03385 [cs]*, Dec. 2015, Accessed: Apr. 03, 2018. [Online]. Available: http://arxiv.org/abs/1512.03385

[81] R. K. Srivastava, K. Greff, and J. Schmidhuber, “Training Very Deep Networks,” in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 2377–2385. Accessed: Dec. 05, 2018. [Online]. Available: http://papers.nips.cc/paper/5850-training-very-deep-networks.pdf

[82] J. Kim, J. K. Lee, and K. M. Lee, "Accurate Image Super-Resolution Using Very Deep Convolutional Networks," *arXiv:1511.04587 [cs]*, Nov. 2015, Accessed: Apr. 03, 2018. [Online]. Available: http://arxiv.org/abs/1511.04587

[83] B. E. Treeby and B. T. Cox, "k-Wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields," *J Biomed Opt*, vol. 15, no. 2, p. 021314, Apr. 2010, doi: 10.1117/1.3360308.

[84] "- k-Wave MATLAB Toolbox." http://www.k-wave.org/documentation/example_pr_2D_tr_circular_sensor.php (accessed Jun. 25, 2018).

[85] A. Dorr, J. G. Sled, and N. Kabani, "Three-dimensional cerebral vasculature of the CBA mouse brain: a magnetic resonance imaging and micro computed tomography study," *Neuroimage*, vol. 35, no. 4, pp. 1409–1423, May 2007, doi: 10.1016/j.neuroimage.2006.12.040.

[86] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale vessel enhancement filtering," in *Medical Image Computing and Computer-Assisted Intervention — MICCAI'98*, vol. 1496, W. M. Wells, A. Colchester, and S. Delp, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 130–137. doi: 10.1007/BFb0056195.

[87] M. Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," *arXiv:1603.04467 [cs]*, Mar. 2016, Accessed: May 04, 2019. [Online]. Available: http://arxiv.org/abs/1603.04467

[88] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.

[89] B. Huang, J. Xia, K. Maslov, and L. V. Wang, "Improving limited-view photoacoustic tomography with an acoustic reflector," *J Biomed Opt*, vol. 18, no. 11, Nov. 2013, doi: 10.1117/1.JBO.18.11.110505.

[90] D. Wu, X. Wang, C. Tao, and X. J. Liu, "Limited-view photoacoustic tomography utilizing backscatterers as virtual transducers," *Appl. Phys. Lett.*, vol. 99, no. 24, p. 244102, Dec. 2011, doi: 10.1063/1.3669512.

[91] C. M. Sandino, N. Dixit, J. Y. Cheng, and S. S. Vasanawala, "Deep convolutional neural networks for accelerated dynamic magnetic resonance imaging," 2017. /paper/Deep-convolutional-neural-networks-for-accelerated-Sandino-Dixit/de12d079e3821ee22586682594d399cbc59d3ff0 (accessed Aug. 03, 2018).

[92] D. Allman, A. Reiter, and M. A. L. Bell, "Photoacoustic Source Detection and Reflection Artifact Removal Enabled by Deep Learning," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1464–1477, Jun. 2018, doi: 10.1109/TMI.2018.2829662.

[93] N. Davoudi, X. L. Deán-Ben, and D. Razansky, "Deep learning optoacoustic tomography with sparse data," *Nature Machine Intelligence*, pp. 1–8, Sep. 2019, doi: 10.1038/s42256-019-0095-3.

[94] S. Antholzer, J. Schwab, and M. Haltmeier, "Deep Learning Versus $^{1\$}$ - Minimization for Compressed Sensing Photoacoustic Tomography," in *2018 IEEE*

International Ultrasonics Symposium (IUS), Oct. 2018, pp. 206–212. doi: 10.1109/ULTSYM.2018.8579737.

[95] H. Lan, K. Zhou, C. Yang, J. Liu, S. Gao, and F. Gao, "Hybrid Neural Network for Photoacoustic Imaging Reconstruction," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2019, pp. 6367–6370. doi: 10.1109/EMBC.2019.8857019.

[96] H. Lan *et al.*, "Ki-GAN: Knowledge Infusion Generative Adversarial Network for Photoacoustic Image Reconstruction In Vivo," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Cham, 2019, pp. 273–281. doi: 10.1007/978-3-030-32239-7_31.

[97] A. Hauptmann *et al.*, "Approximate k-Space Models and Deep Learning for Fast Photoacoustic Reconstruction," in *Machine Learning for Medical Image Reconstruction*, Cham, 2018, pp. 103–111. doi: 10.1007/978-3-030-00129-2_12.

[98] J. Adler and O. Öktem, "Solving ill-posed inverse problems using iterative deep neural networks," *Inverse Problems*, vol. 33, no. 12, p. 124007, 2017, doi: 10.1088/1361-6420/aa9581.

[99] J. Schwab, S. Antholzer, and M. Haltmeier, "Learned backprojection for sparse and limited view photoacoustic tomography," in *Photons Plus Ultrasound: Imaging and Sensing 2019*, Feb. 2019, vol. 10878, p. 1087837. doi: 10.1117/12.2508438.

[100] A. Budai, R. Bock, A. Maier, J. Hornegger, and G. Michelson, "Robust Vessel Segmentation in Fundus Images," *International Journal of Biomedical Imaging*, 2013. https://www.hindawi.com/journals/ijbi/2013/154860/ (accessed Dec. 16, 2019).

[101] "Public Lung Image Database." http://www.via.cornell.edu/lungdb.html (accessed Dec. 16, 2019).

[102] B. E. Treeby and B. T. Cox, "k-Wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields," *J Biomed Opt*, vol. 15, no. 2, p. 021314, Apr. 2010, doi: 10.1117/1.3360308.

[103] A. Hauptmann and B. Cox, "Deep Learning in Photoacoustic Tomography: Current approaches and future directions," *arXiv:2009.07608 [cs, eess]*, Sep. 2020, Accessed: Nov. 18, 2020. [Online]. Available: http://arxiv.org/abs/2009.07608

[104] C. Yang, H. Lan, F. Gao, and F. Gao, "Deep learning for photoacoustic imaging: a survey," *arXiv:2008.04221 [cs, eess]*, Nov. 2020, Accessed: Nov. 18, 2020. [Online]. Available: http://arxiv.org/abs/2008.04221

[105] H. Shan *et al.*, "Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction," *Nature Machine Intelligence*, vol. 1, no. 6, p. 269, Jun. 2019, doi: 10.1038/s42256-019-0057-9.

[106] C. M. Hyun, H. P. Kim, S. M. Lee, S. Lee, and J. K. Seo, "Deep learning for undersampled MRI reconstruction," *Phys. Med. Biol.*, vol. 63, no. 13, p. 135007, Jun. 2018, doi: 10.1088/1361-6560/aac71a.

[107] S. Guan, A. A. Khan, S. Sikdar, and P. V. Chitnis, "Limited-View and Sparse Photoacoustic Tomography for Neuroimaging with Deep Learning," *Scientific Reports*, vol. 10, no. 1, Art. no. 1, May 2020, doi: 10.1038/s41598-020-65235-2.

[108] M. Kim, G.-S. Jeng, I. Pelivanov, and M. O'Donnell, "Deep-Learning Image Reconstruction for Real-Time Photoacoustic System," *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, pp. 3379–3390, Nov. 2020, doi: 10.1109/TMI.2020.2993835.
[109] F. Yu, V. Koltun, and T. Funkhouser, "Dilated Residual Networks," *arXiv:1705.09914 [cs]*, May 2017, Accessed: Dec. 09, 2020. [Online]. Available: http://arxiv.org/abs/1705.09914
[110] Z. Wang and S. Ji, "Smoothed Dilated Convolutions for Improved Dense Prediction," *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2486–2495, Jul. 2018, doi: 10.1145/3219819.3219944.
[111] F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," *arXiv:1511.07122 [cs]*, Apr. 2016, Accessed: Nov. 18, 2020. [Online]. Available: http://arxiv.org/abs/1511.07122
[112] P. Wang *et al.*, "Understanding Convolution for Semantic Segmentation," *arXiv:1702.08502 [cs]*, May 2018, Accessed: Dec. 09, 2020. [Online]. Available: http://arxiv.org/abs/1702.08502
[113] H. Zhou *et al.*, "Multi-Scale Dilated Convolution Neural Network for Image Artifact Correction of Limited-Angle Tomography," *IEEE Access*, vol. 8, pp. 1567–1576, 2020, doi: 10.1109/ACCESS.2019.2962071.
[114] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, Feb. 2001, doi: 10.1109/83.902291.
[115] A. Badano *et al.*, "Evaluation of Digital Breast Tomosynthesis as Replacement of Full-Field Digital Mammography Using an In Silico Imaging Trial," *JAMA Netw Open*, vol. 1, no. 7, p. e185474, Nov. 2018, doi: 10.1001/jamanetworkopen.2018.5474.
[116] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, Nov. 2003, vol. 2, pp. 1398-1402 Vol.2. doi: 10.1109/ACSSC.2003.1292216.
[117] M. Xu, Y. Xu, and L. V. Wang, "Time-domain reconstruction algorithms and numerical simulations for thermoacoustic tomography in various geometries," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 9, pp. 1086–1099, Sep. 2003, doi: 10.1109/TBME.2003.816081.
[118] G. Paltauf, J. A. Viator, S. A. Prahl, and S. L. Jacques, "Iterative reconstruction algorithm for optoacoustic imaging," *The Journal of the Acoustical Society of America*, vol. 112, no. 4, pp. 1536–1544, Sep. 2002, doi: 10.1121/1.1501898.
[119] B. Baumann, M. Wolff, B. Kost, and H. Groninga, "Finite element calculation of photoacoustic signals," *Appl. Opt., AO*, vol. 46, no. 7, pp. 1120–1125, Mar. 2007, doi: 10.1364/AO.46.001120.
[120] W. Xia *et al.*, "An optimized ultrasound detector for photoacoustic breast tomography," *Med Phys*, vol. 40, no. 3, p. 032901, Mar. 2013, doi: 10.1118/1.4792462.

[121] M. Raissi, P. Perdikaris, and G. E. Karniadakis, “Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations,” *Journal of Computational Physics*, vol. 378, pp. 686–707, Feb. 2019, doi: 10.1016/j.jcp.2018.10.045.
[122] D. Greenfeld, M. Galun, R. Basri, I. Yavneh, and R. Kimmel, “Learning to Optimize Multigrid PDE Solvers,” in *International Conference on Machine Learning*, May 2019, pp. 2415–2423. Accessed: May 22, 2021. [Online]. Available: http://proceedings.mlr.press/v97/greenfeld19a.html
[123] Y. Khoo, J. Lu, and L. Ying, “Solving parametric PDE problems with artificial neural networks,” *Eur. J. Appl. Math*, vol. 32, no. 3, pp. 421–435, Jun. 2021, doi: 10.1017/S0956792520000182.
[124] W. E and B. Yu, “The Deep Ritz method: A deep learning-based numerical algorithm for solving variational problems,” *arXiv:1710.00211 [cs, stat]*, Sep. 2017, Accessed: May 22, 2021. [Online]. Available: http://arxiv.org/abs/1710.00211
[125] L. Lu, P. Jin, and G. E. Karniadakis, “DeepONet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators,” *arXiv:1910.03193 [cs, stat]*, Apr. 2020, Accessed: May 22, 2021. [Online]. Available: http://arxiv.org/abs/1910.03193
[126] Z. Li *et al.*, “Fourier Neural Operator for Parametric Partial Differential Equations,” *arXiv:2010.08895 [cs, math]*, Oct. 2020, Accessed: Dec. 29, 2020. [Online]. Available: http://arxiv.org/abs/2010.08895
[127] E. Tadmor, “A review of numerical methods for nonlinear partial differential equations,” *Bull. Amer. Math. Soc.*, vol. 49, no. 4, pp. 507–554, 2012, doi: 10.1090/S0273-0979-2012-01379-4.
[128] B. E. Treeby and J. Pan, “A practical examination of the errors arising in the direct collocation boundary element method for acoustic scattering,” *Engineering Analysis with Boundary Elements*, vol. 33, no. 11, pp. 1302–1315, Nov. 2009, doi: 10.1016/j.enganabound.2009.06.005.
[129] T. D. Mast, L. P. Souriau, D.-L. D. Liu, M. Tabei, A. I. Nachman, and R. C. Waag, “A k-space method for large-scale models of wave propagation in tissue,” *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 48, no. 2, pp. 341–354, Mar. 2001, doi: 10.1109/58.911717.
[130] B. E. Treeby, J. Jaros, A. P. Rendell, and B. T. Cox, “Modeling nonlinear ultrasound propagation in heterogeneous media with power law absorption using a k-space pseudospectral method,” *J Acoust Soc Am*, vol. 131, no. 6, pp. 4324–4336, Jun. 2012, doi: 10.1121/1.4712021.
[131] Z. Akkus *et al.*, “A Survey of Deep-Learning Applications in Ultrasound: Artificial Intelligence–Powered Ultrasound for Improving Clinical Workflow,” *Journal of the American College of Radiology*, vol. 16, no. 9, Part B, pp. 1318–1328, Sep. 2019, doi: 10.1016/j.jacr.2019.06.004.
[132] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, “What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use,” in *Machine Learning for Healthcare Conference*, Oct. 2019, pp. 359–380. Accessed:

Aug. 04, 2021. [Online]. Available: http://proceedings.mlr.press/v106/tonekaboni19a.html

[133] C. Cai, K. Deng, C. Ma, and J. Luo, "End-to-end deep neural network for optical inversion in quantitative photoacoustic imaging," *Opt. Lett., OL*, vol. 43, no. 12, pp. 2752–2755, Jun. 2018, doi: 10.1364/OL.43.002752.

BIOGRAPHY

Steven Guan graduated from Fairfax High School, Fairfax, Virginia, in 1983. She received her Bachelor of Arts from George Mason University in 1987. She was employed as a teacher in Fairfax County for two years and received her Master of Arts in English from George Mason University in 1987.