Solving Mathematical Problems by Deep Learning: Inverse Problems

Gitta Kutyniok

(Technische Universität Berlin and University of Tromsø)

2019 Woudschoten Conference
Zeist, The Netherlands, October 9–11, 2019

The 21st Century

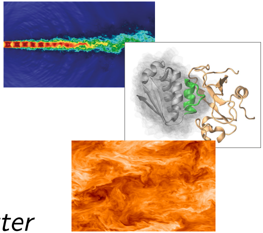
Various technological advances in the 21st century are only possible through *integrated mathematical modeling, simulation, and optimization*.

The 21st Century

Various technological advances in the 21st century are only possible through *integrated mathematical modeling, simulation, and optimization*.

Further Examples:

- Turbines
  $\rightsquigarrow$ *Adjoint based jet-noise minimization*
- Atomistic molecular dynamics
  $\rightsquigarrow$ *Simulations with ultralong timescales*
- Star formation
  $\rightsquigarrow$ *Understanding of turbulent accretion of matter*

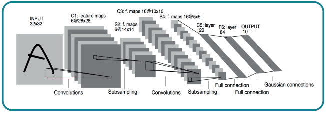


There is a pressing need to go beyond pure modeling, simulation, and optimization approaches!

The Data Science Side: Impact of Deep Learning

Self-Driving Cars

Surveillance



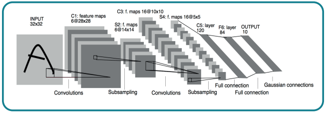
Legal Issues

Health Care

The Data Science Side: Impact of Deep Learning

Self-Driving Cars

Surveillance



Legal Issues

Health Care

Very few theoretical results explaining their success!

Deep Learning = Alchemy?

SIAM NEWS MAY 2017

Research | May 01, 2017

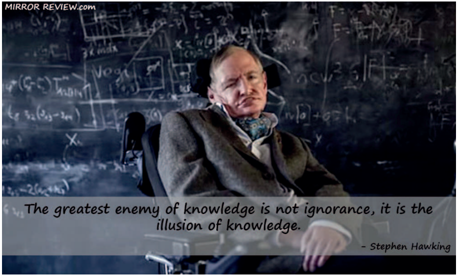
Deep, Deep Trouble

Deep Learning's Impact on Image Processing, Mathematics, and Humanity

By Michael Elad

„Ali Rahimi, a researcher in artificial intelligence (AI) at Google in San Francisco, California, took a swipe at his field last December—and received a 40-second ovation for it. Speaking at an AI conference, Rahimi charged that **machine learning algorithms, in which computers learn through trial and error, have become a form of „alchemy."** Researchers, he said, **do not know why some algorithms work and others don't, nor do they have rigorous criteria for choosing one AI architecture over another....**"

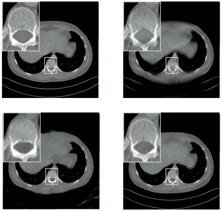
Science, May 2018



Impact of Deep Learning on Mathematics

Some Examples:

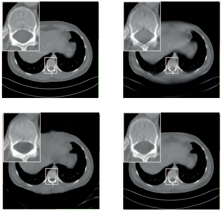
- Inverse Problems
  - ⇝ *Image denoising (Burger, Schuler, Harmeling; 2012)*
  - ⇝ *Superresolution (Klatzer, Soukup, Kobler, Hammernik, Pock; 2017)*
  - ⇝ *Limited-angle tomography (Bubba, K, Lassas, März, Samek, Siltanen, Srinivan; 2018)*
  - ⇝ *Edge detection (Andrade-Loarca, K, Öktem, Petersen; 2019)*



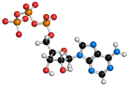
Impact of Deep Learning on Mathematics

Some Examples:

- Inverse Problems
  - ⇝ *Image denoising (Burger, Schuler, Harmeling; 2012)*
  - ⇝ *Superresolution (Klatzer, Soukup, Kobler, Hammernik, Pock; 2017)*
  - ⇝ *Limited-angle tomography (Bubba, K, Lassas, März, Samek, Siltanen, Srinivan; 2018)*
  - ⇝ *Edge detection (Andrade-Loarca, K, Öktem, Petersen; 2019)*



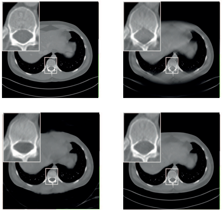
- Numerical Analysis of Partial Differential Equations
  - ⇝ *Schrödinger equation (Rupp, Tkatchenko, Müller, von Lilienfeld; 2012 –)*
  - ⇝ *Black-Scholes PDEs (Grohs, Hornung, Jentzen,von Wurstemberger; 2018)*
  - ⇝ *Parametric PDEs (Schwab, Zech; 2018)*
  - ⇝ *Parametric PDEs (K, Petersen, Raslan, Schneider; 2019)*



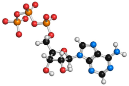
Impact of Deep Learning on Mathematics

Some Examples:

- Inverse Problems
  - ⇝ *Image denoising (Burger, Schuler, Harmeling; 2012)*
  - ⇝ *Superresolution (Klatzer, Soukup, Kobler, Hammernik, Pock; 2017)*
  - ⇝ *Limited-angle tomography (Bubba, K, Lassas, März, Samek, Siltanen, Srinivan; 2018)*
  - ⇝ *Edge detection (Andrade-Loarca, K, Öktem, Petersen; 2019)*



- Numerical Analysis of Partial Differential Equations
  - ⇝ *Schrödinger equation (Rupp, Tkatchenko, Müller, von Lilienfeld; 2012 –)*
  - ⇝ *Black-Scholes PDEs (Grohs, Hornung, Jentzen,von Wurstemberger; 2018)*
  - ⇝ *Parametric PDEs (Schwab, Zech; 2018)*
  - ⇝ *Parametric PDEs (K, Petersen, Raslan, Schneider; 2019)*

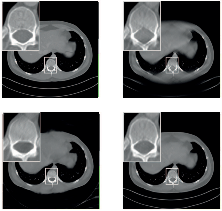


- Modelling
  - ⇝ *Learning equations from data (Sahoo, Lampert, Martius; 2018)*

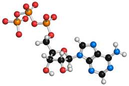
Impact of Deep Learning on Mathematics

Some Examples:

- Inverse Problems
  - $\rightsquigarrow$ *Image denoising (Burger, Schuler, Harmeling; 2012)*
  - $\rightsquigarrow$ *Superresolution (Klatzer, Soukup, Kobler, Hammernik, Pock; 2017)*
  - $\rightsquigarrow$ *Limited-angle tomography (Bubba, K, Lassas, März, Samek, Siltanen, Srinivan; 2018)*
  - $\rightsquigarrow$ *Edge detection (Andrade-Loarca, K, Öktem, Petersen; 2019)*



- Numerical Analysis of Partial Differential Equations
  - $\rightsquigarrow$ *Schrödinger equation (Rupp, Tkatchenko, Müller, von Lilienfeld; 2012 –)*
  - $\rightsquigarrow$ *Black-Scholes PDEs (Grohs, Hornung, Jentzen,von Wurstemberger; 2018)*
  - $\rightsquigarrow$ *Parametric PDEs (Schwab, Zech; 2018)*
  - $\rightsquigarrow$ *Parametric PDEs (K, Petersen, Raslan, Schneider; 2019)*



- Modelling
  - $\rightsquigarrow$ *Learning equations from data (Sahoo, Lampert, Martius; 2018)*

Data-Driven Versus Model-Based Approaches?

Problems, Viewpoints and Solution Strategies:

- Detect structural components in data sets
  Deep neural networks, ...
- Insert physical information in machine learning algorithm
  Machine learning with physical constraints, ...
- Learn parameters from given data sets
  Parametric differential equations, ...
- Study simulation generated data in search of underlying laws
  Data analysis on simulation data, ...

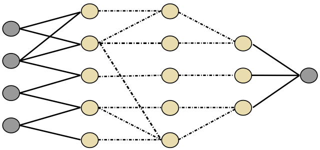
Optimal balancing of
data-driven and model-based approaches!

Mathematics of Deep Neural Networks

The Mathematics of Deep Neural Networks

Definition:

Assume the following notions:



- $d \in \mathbb{N}$: Dimension of input layer.
- L: Number of layers.
- N: Number of neurons.
- $\rho : \mathbb{R} \to \mathbb{R}$: (Non-linear) function called *activation function*.
- $T_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$, $\ell = 1, \ldots, L$: Affine linear maps.

Then $\Phi : \mathbb{R}^d \to \mathbb{R}^{N_L}$ given by

$$\Phi(x) = T_L \rho(T_{L-1} \rho(\ldots \rho(T_1(x)))), \quad x \in \mathbb{R}^d,$$

is called *(deep) neural network (DNN)*.

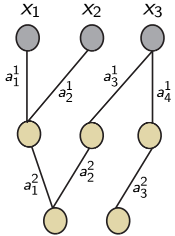
Affine Linear Maps and Weights

Remark: The affine linear map T_ℓ is defined by a matrix $A_\ell \in \mathbb{R}^{N_{\ell-1} \times N_\ell}$ and an affine part $b_\ell \in \mathbb{R}^{N_\ell}$ via

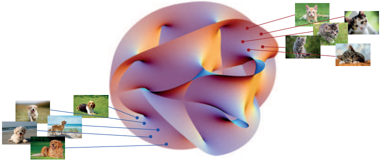
$$T_\ell(x) = A_\ell x + b_\ell.$$

$$A_1 = \begin{pmatrix} a_1^1 & a_2^1 & 0 \\ 0 & 0 & a_3^1 \\ 0 & 0 & a_4^1 \end{pmatrix}$$

$$A_2 = \begin{pmatrix} a_1^2 & a_2^2 & 0 \\ 0 & 0 & a_3^2 \end{pmatrix}$$



Training of Deep Neural Networks



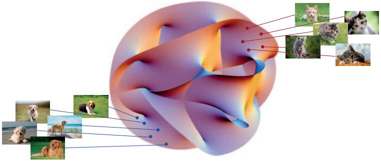
High-Level Set Up:

- Samples $(x_i, f(x_i))_{i=1}^m$ of a function such as $f : \mathcal{M} \to \{1, 2, \ldots, K\}$.

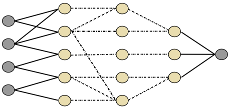
Training of Deep Neural Networks

High-Level Set Up:

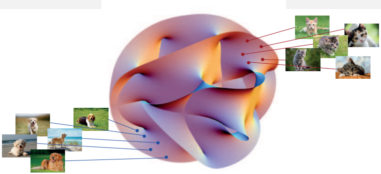
- Samples $(x_i, f(x_i))_{i=1}^m$ of a function such as $f : \mathcal{M} \to \{1, 2, \dots, K\}$.



- Select an architecture of a deep neural network, i.e., a choice of d, L, $(N_\ell)_{\ell=1}^L$, and ρ.
  Sometimes selected entries of the matrices $(A_\ell)_{\ell=1}^L$, i.e., weights, are set to zero at this point.

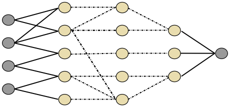


Training of Deep Neural Networks



High-Level Set Up:

- Samples $(x_i, f(x_i))_{i=1}^m$ of a function such as $f : \mathcal{M} \to \{1, 2, \ldots, K\}$.



- Select an architecture of a deep neural network, i.e., a choice of d, L, $(N_\ell)_{\ell=1}^L$, and ρ.
  Sometimes selected entries of the matrices $(A_\ell)_{\ell=1}^L$, i.e., weights, are set to zero at this point.
- Learn the affine-linear functions $(T_\ell)_{\ell=1}^L = (A_\ell \cdot + b_\ell)_{\ell=1}^L$ by

$$\min_{(A_\ell, b_\ell)_\ell} \sum_{i=1}^m \mathcal{L}(\Phi_{(A_\ell,b_\ell)_\ell}(x_i), f(x_i)) + \lambda \mathcal{R}((A_\ell, b_\ell)_\ell)$$

yielding the network $\Phi_{(A_\ell,b_\ell)_\ell} : \mathbb{R}^d \to \mathbb{R}^{N_L}$,

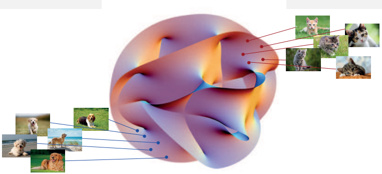
$$\Phi_{(A_\ell,b_\ell)_\ell}(x) = T_L \rho(T_{L-1} \rho(\ldots \rho(T_1(x)))).$$

This is often done by stochastic gradient descent.

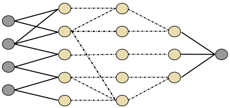
Training of Deep Neural Networks

High-Level Set Up:

- Samples $(x_i, f(x_i))_{i=1}^m$ of a function such as $f : \mathcal{M} \to \{1, 2, \ldots, K\}$.



- Select an architecture of a deep neural network, i.e., a choice of d, L, $(N_\ell)_{\ell=1}^L$, and ρ.
  Sometimes selected entries of the matrices $(A_\ell)_{\ell=1}^L$, i.e., weights, are set to zero at this point.



- Learn the affine-linear functions $(T_\ell)_{\ell=1}^L = (A_\ell \cdot + b_\ell)_{\ell=1}^L$ by

$$\min_{(A_\ell, b_\ell)_\ell} \sum_{i=1}^m \mathcal{L}(\Phi_{(A_\ell,b_\ell)_\ell}(x_i), f(x_i)) + \lambda \mathcal{R}((A_\ell, b_\ell)_\ell)$$

yielding the network $\Phi_{(A_\ell,b_\ell)_\ell} : \mathbb{R}^d \to \mathbb{R}^{N_L}$,

$$\Phi_{(A_\ell,b_\ell)_\ell}(x) = T_L \rho(T_{L-1} \rho(\ldots \rho(T_1(x)))).$$

This is often done by stochastic gradient descent.

Goal: $\Phi_{(A_\ell,b_\ell)_\ell} \approx f$

Fundamental Questions concerning Deep Neural Networks

- *Expressivity:*
  - How powerful is the network architecture?
  - Can it indeed represent the correct functions?

$\rightsquigarrow$ *Applied Harmonic Analysis, Approximation Theory, ...*

Fundamental Questions concerning Deep Neural Networks

- *Expressivity:*
  - How powerful is the network architecture?
  - Can it indeed represent the correct functions?

$\rightsquigarrow$ *Applied Harmonic Analysis, Approximation Theory, ...*

- *Learning:*
  - Why does the current learning algorithm produce anything reasonable?
  - What are good starting values?

$\rightsquigarrow$ *Differential Geometry, Optimal Control, Optimization, ...*

Fundamental Questions concerning Deep Neural Networks

- *Expressivity:*
  - How powerful is the network architecture?
  - Can it indeed represent the correct functions?

$\rightsquigarrow$ *Applied Harmonic Analysis, Approximation Theory, ...*

- *Learning:*
  - Why does the current learning algorithm produce anything reasonable?
  - What are good starting values?

$\rightsquigarrow$ *Differential Geometry, Optimal Control, Optimization, ...*

- *Generalization:*
  - Why do deep neural networks perform that well on data sets, which do not belong to the input-output pairs from a training set?
  - What impact has the depth of the network?

$\rightsquigarrow$ *Learning Theory, Optimization, Statistics, ...*

Fundamental Questions concerning Deep Neural Networks

- *Expressivity:*
  - How powerful is the network architecture?
  - Can it indeed represent the correct functions?

$\rightsquigarrow$ *Applied Harmonic Analysis, Approximation Theory, ...*

- *Learning:*
  - Why does the current learning algorithm produce anything reasonable?
  - What are good starting values?

$\rightsquigarrow$ *Differential Geometry, Optimal Control, Optimization, ...*

- *Generalization:*
  - Why do deep neural networks perform that well on data sets, which do not belong to the input-output pairs from a training set?
  - What impact has the depth of the network?

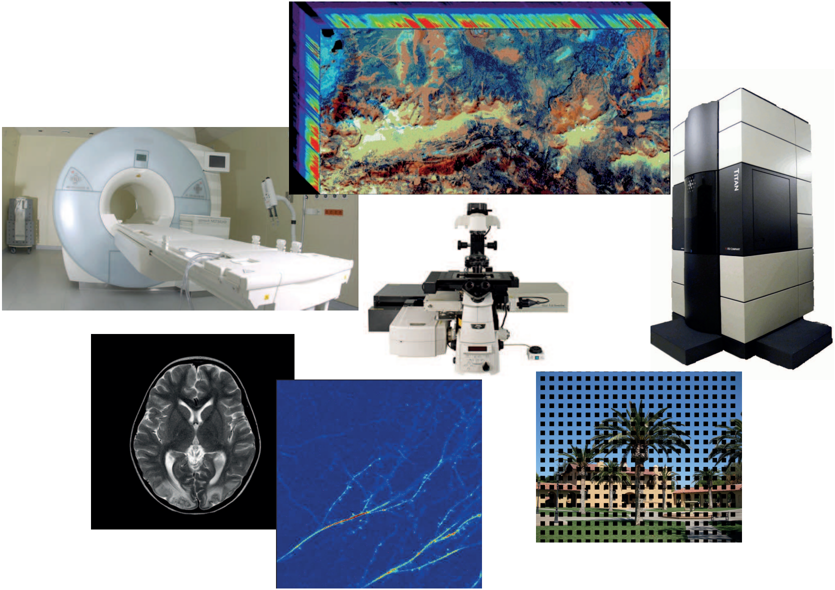
$\rightsquigarrow$ *Learning Theory, Optimization, Statistics, ...*

- *Interpretability:*
  - Why did a trained deep neural network reach a certain decision?
  - Which components of the input do contribute most?

$\rightsquigarrow$ *Information Theory, Uncertainty Quantification, ...*

Inverse Problems in Imaging

Modern Imaging Science



Solving Inverse Problems

Tikhonov Regularization:
Given an (ill-posed) inverse problem

$$Kf = g, \qquad \text{where} \quad K : X \to Y,$$

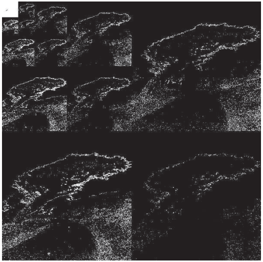
an approximate solution $f^\alpha \in X$, $\alpha > 0$, can be determined by

$$f^\alpha := \operatorname{argmin}_{f \in X} \Big[\underbrace{\|Kf - g\|^2}_{\text{Data fidelity term}} + \alpha \cdot \underbrace{\mathcal{P}(f)}_{\text{Penalty term}} \Big].$$

Penalty Term: The penalty term $\mathcal{P}$

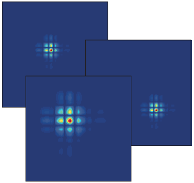
- ensures continuous dependence on the data,
- incorporates properties of the solution.

The World is Compressible!



Wavelet Transform (JPEG2000):

$$f \mapsto (\langle f, \psi_{j,m} \rangle)_{j,m}.$$



Definition: For a wavelet $\psi \in L^2(\mathbb{R}^2)$, a wavelet system is defined by

$$\{\psi_{j,m} : j \in \mathbb{Z}, m \in \mathbb{Z}^2\}, \quad \text{where } \psi_{j,m}(x) := 2^j \psi(2^j x - m).$$

Sparsity

Paradigm:

For each class of data, there exists a sparsifying system!

Sparsity

Paradigm:

For each class of data, there exists a sparsifying system!

Two Viewpoints of 'Sparsifying System':

Let $\mathcal{C} \subseteq \mathcal{H}$ and $(\psi_\lambda)_\lambda \subseteq \mathcal{H}$.

- Decay of Coefficients. Consider the decay for $n \to \infty$ of the sorted sequence of coefficients

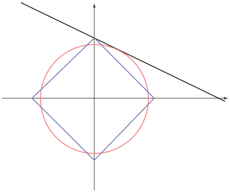
$$(|\langle x, \psi_{\lambda_n} \rangle|)_n \quad \text{for all } x \in \mathcal{C}.$$

- Approximation Properties. Consider the decay for $N \to \infty$ of the error of best N-term approximation, i.e.,

$$\inf_{\#\Lambda_N = N, (c_\lambda)_\lambda} \left\| x - \sum_{\lambda \in \Lambda_N} c_\lambda \psi_\lambda \right\| \quad \text{for all } x \in \mathcal{C}.$$

How to Penalize Non-Sparsity?

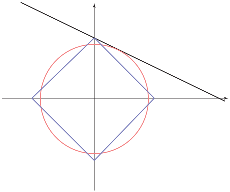
Intuition:



$\rightsquigarrow$ *Use the ℓ_1 norm!*

How to Penalize Non-Sparsity?

Intuition:



$\rightsquigarrow$ *Use the ℓ_1 norm!*

Sparse Regularization:

Solve an ill-posed inverse problem $Kf = g$ by

$$f^\alpha := \mathrm{argmin}_f \Big[\underbrace{\|Kf - g\|^2}_{\text{Data fidelity term}} + \alpha \cdot \underbrace{\|(\langle f, \psi_{j,m} \rangle)_{j,m}\|_1}_{\text{Penalty term}} \Big].$$

Sparsity-Based Approaches to Inverse Problems

Compressed Sensing (Candès, Romberg, Tao and Donoho; 2006) :

- Goal: Solve an underdetermined linear problem

$$y = Ax, \quad A \text{ an } n \times N\text{-matrix with } n \ll N,$$

for a solution $x \in \mathbb{R}^N$ admitting a sparsifying system $(\psi_\lambda)_\lambda$.

- Approach: Recover x by the ℓ_1-analysis minimization problem

$$\min_{\tilde{x}} \|(\langle \tilde{x}, \psi_\lambda \rangle)_\lambda\|_1 \text{ subject to } y = A\tilde{x}$$

Some Earlier Footprints in Inverse Problems:

- Donoho (1995): Wavelet-Vaguelette decomposition.
- Chambolle, DeVore, Lee, Lucier (1998): Penalty on the Besov norm.
- Daubechies, Defries, De Mol (2004): General sparsity constraints.
- ...

Shearlets come into Play

Mathematical Model for Images

Key Observation:

Images are governed by edge-like structures!

Mathematical Model for Images

Key Observation:

Images are governed by edge-like structures!

Definition (Donoho; 2001):

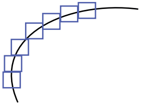
Let $\nu > 0$. We then define the class of *cartoon-like functions* by

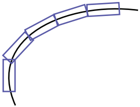
$$\mathcal{E}^2(\mathbb{R}^2) = \{f \in L^2(\mathbb{R}^2) : f = f_1 + \chi_B f_2\},$$

where $B \subset [0,1]^2$ with $\partial B \in C^2$, and the functions f_1 and f_2 satisfy $f_1, f_2 \in C_0^2([0,1]^2)$, $\|f_1\|_{C^2}, \|f_2\|_{C^2}, \|\partial B\|_{C^2} < \nu$.

Key Ideas of the Shearlet Construction

Wavelet versus Shearlet Approximation:



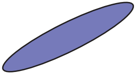


Key Ideas of the Shearlet Construction

Wavelet versus Shearlet Approximation:

Parabolic scaling ('width $\approx$ length^2'):

$$A_{2^j} = \begin{pmatrix} 2^j & 0 \\ 0 & 2^{j/2} \end{pmatrix}, \quad j \in \mathbb{Z}.$$



Orientation via shearing:

$$S_k = \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}, \quad k \in \mathbb{Z}.$$

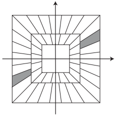
Advantage:

- Shearing leaves the digital grid $\mathbb{Z}^2$ invariant.
- Uniform theory for the continuum and digital situation.

(Cone-adapted) Discrete Shearlet Systems

Definition (K, Labate; 2006):

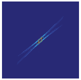
The (cone-adapted) discrete shearlet system $\mathcal{SH}(c;\phi,\psi,\tilde{\psi})$, $c > 0$, generated by $\phi \in L^2(\mathbb{R}^2)$ and $\psi, \tilde{\psi} \in L^2(\mathbb{R}^2)$ is the union of



$$\{\phi(\cdot - cm) : m \in \mathbb{Z}^2\},$$

$$\{2^{3j/4}\psi(S_k A_{2^j} \cdot - cm) : j \geq 0, |k| \leq \lceil 2^{j/2} \rceil, m \in \mathbb{Z}^2\},$$

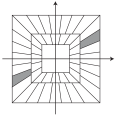
$$\{2^{3j/4}\tilde{\psi}(\tilde{S}_k \tilde{A}_{2^j} \cdot - cm) : j \geq 0, |k| \leq \lceil 2^{j/2} \rceil, m \in \mathbb{Z}^2\}.$$



(Cone-adapted) Discrete Shearlet Systems

Definition (K, Labate; 2006):

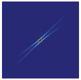
The (cone-adapted) discrete shearlet system $\mathcal{SH}(c;\phi,\psi,\tilde{\psi})$, $c>0$, generated by $\phi \in L^2(\mathbb{R}^2)$ and $\psi,\tilde{\psi} \in L^2(\mathbb{R}^2)$ is the union of



$$\{\phi(\cdot - cm) : m \in \mathbb{Z}^2\},$$

$$\{2^{3j/4}\psi(S_k A_{2^j}\,\cdot\, - cm) : j \geq 0, |k| \leq \lceil 2^{j/2}\rceil, m \in \mathbb{Z}^2\},$$

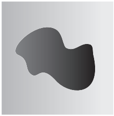
$$\{2^{3j/4}\tilde{\psi}(\tilde{S}_k \tilde{A}_{2^j}\,\cdot\, - cm) : j \geq 0, |k| \leq \lceil 2^{j/2}\rceil, m \in \mathbb{Z}^2\}.$$



Theorem (K, Lim; 2011):

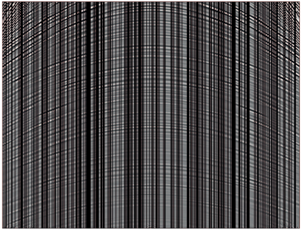
Let $\phi,\psi,\tilde{\psi} \in L^2(\mathbb{R}^2)$ be compactly supported, and let $\hat{\psi}$, $\hat{\tilde{\psi}}$ satisfy certain decay condition. Then $\mathcal{SH}(\phi,\psi,\tilde{\psi})$ provides an optimally sparse approximation of $f \in \mathcal{E}^2(\mathbb{R}^2)$, i.e.,

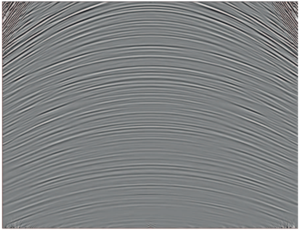
$$\|f - f_N\|_2 \leq C \cdot N^{-1} \cdot (\log N)^{3/2}, \quad N \to \infty.$$



Applications

Inpainting:

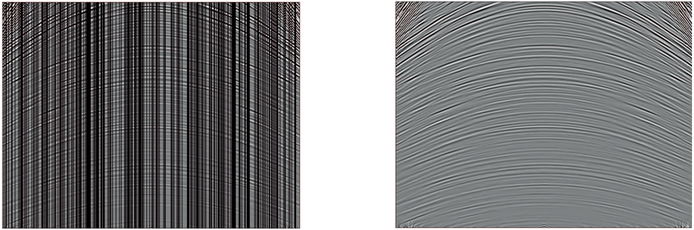




(Source: K, Lim; 2012)

Applications

Inpainting:



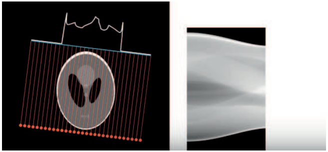
(Source: K, Lim; 2012)

2D&3D (parallelized) Fast Shearlet Transform (`www.ShearLab.org`):

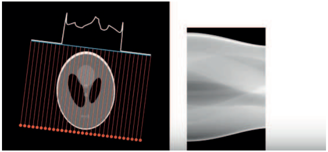
- Matlab *(K, Lim, Reisenhofer; 2013)*
- Julia *(Loarca; 2017)*
- Python *(Look; 2018)*
- Tensorflow *(Loarca; 2019)*

Mathematical Modeling Reaches a Barrier

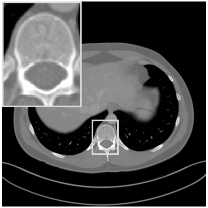
Computed Tomography (CT)

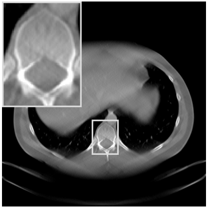


Computed Tomography (CT)



Problem with Limited-Angle Tomography:



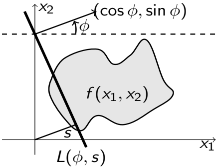


The data is too complex for mathematical modeling!

Limited Angle-(Computed) Tomography

A CT scanner samples the *Radon transform*

$$\mathcal{R}f(\phi, s) = \int_{L(\phi,s)} f(x)dS(x),$$

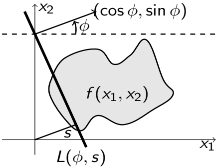


for $L(\phi, s) = \{x \in \mathbb{R}^2 : x_1 \cos(\phi) + x_2 \sin(\phi) = s\}$,
$\phi \in [-\pi/2, \pi/2)$, and $s \in \mathbb{R}$.

Limited Angle-(Computed) Tomography

A CT scanner samples the *Radon transform*

$$\mathcal{R}f(\phi, s) = \int_{L(\phi,s)} f(x)dS(x),$$



for $L(\phi, s) = \{x \in \mathbb{R}^2 : x_1 \cos(\phi) + x_2 \sin(\phi) = s\}$,
$\phi \in [-\pi/2, \pi/2)$, and $s \in \mathbb{R}$.

Challenging inverse problem if $\mathcal{R}f(\cdot, s)$ is only sampled on $[-\phi, \phi] \subset [-\pi/2, \pi/2)$.

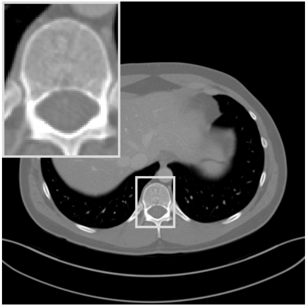
Applications: Dental CT, breast tomosynthesis, electron tomography,...

Model-Based Approaches Fail

Sparse Regularization:

$$\operatorname{argmin}_f \Big[\underbrace{\|\mathcal{R}f - g\|^2}_{\text{Data fidelity term}} + \alpha \cdot \underbrace{\|(\langle f, \psi_{j,k,m} \rangle)_{j,k,m}\|_1}_{\text{Penalty term}} \Big].$$

Clinical Data:



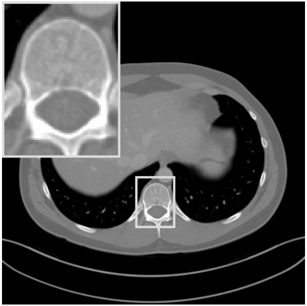
Original Image

Model-Based Approaches Fail

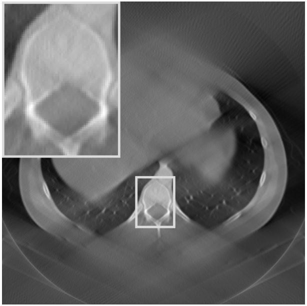
Sparse Regularization:

$$\operatorname{argmin}_f \Big[\underbrace{\|\mathcal{R}f - g\|^2}_{\text{Data fidelity term}} + \alpha \cdot \underbrace{\|(\langle f, \psi_{j,k,m}\rangle)_{j,k,m}\|_1}_{\text{Penalty term}} \Big].$$

Clinical Data:



Original Image



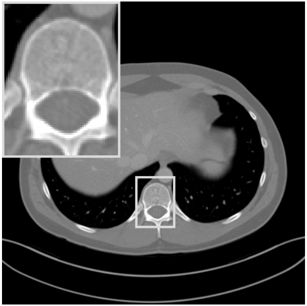
Filtered Backprojection

Model-Based Approaches Fail

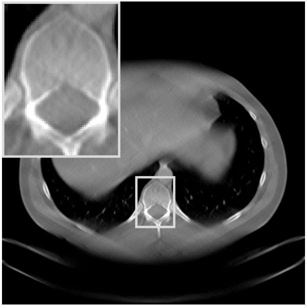
Sparse Regularization:

$$\operatorname{argmin}_f \left[\underbrace{\|\mathcal{R}f - g\|^2}_{\text{Data fidelity term}} + \alpha \cdot \underbrace{\|(\langle f, \psi_{j,k,m} \rangle)_{j,k,m}\|_1}_{\text{Penalty term}} \right].$$

Clinical Data:



Original Image



Sparse Regularization with Shearlets

Let's bring Deep Learning into the Game

Solving Inverse Problems by Deep Learning

Setup:
Given N training samples $(f_i, g_i)_{i=1}^N$ following the forward model

$$g_i = Kf_i + \eta.$$

Goal:

- Determine a reconstruction operator $\mathcal{T}_\theta$ such that

$$g = Kf + \eta \quad \Longrightarrow \quad \mathcal{T}_\theta(g) \approx f.$$

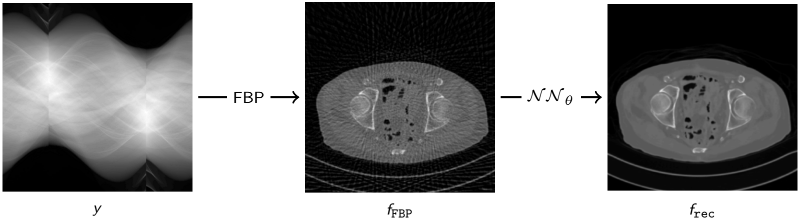
- $\mathcal{T}_\theta$ is parametrized by $\theta \in \mathbb{R}^p$ and learned from training data.

Evaluation:
Evaluate the quality of $\mathcal{T}_\theta$ by testing on the test data $(f_i, g_i)_{i=N+1}^K$ following the forward model.

Denoising Direct Inversions

Denoising Direct Inversion (Ye et al.; 2016), (Unser et. al.; 2017), ...:

- Idea: Direct inversion with filtered backprojection, train CNN to remove noise.
- Illustration:

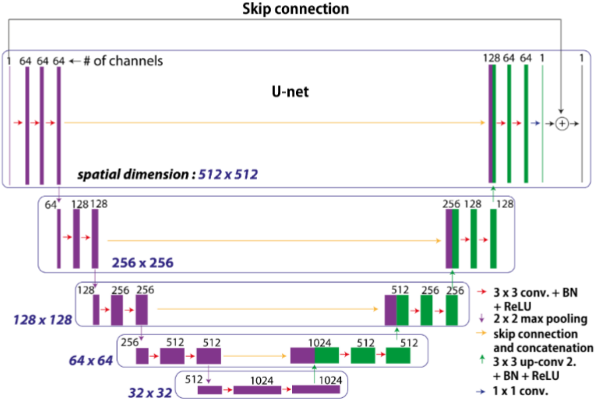


Inversion & denoising $\rightsquigarrow$ Simple, ad-hoc approach to inverse problems

- Intuition:
  - CNN learns structured noise/artifacts.
  - Rationale: Without taking FBP, CNN needs to learn physics of CT.

Denoising Direct Inversions - The CNN Architecture

- U-Net architecture, originally used for segmentation (Ronneberger et al.; 2015)
- Based on fully-convolutional networks (Long et al.; 2014)
- Encoder-Decoder CNN with skip-connections



Other Deep Learning Approaches to Inverse Problems

Tikhonov Regularization:

$$\mathrm{argmin}_f \left[\|Kf - g\|^2 + \ \alpha \cdot \mathcal{P}(f) \right]$$

Plug-and-play with CNN-denoising [Bouman et al.,2013], [Elad et al.,2016], ...

- Iterative solvers such as Douglas-Rachford or ADMM contain a denoising step.
- Replace this step by a trained CNN.

Learned Iterative Schemes [Pock et. al.,2017], [Adler et al.,2017], ...

- Iterative solvers such as ADMM or Primal-Dual contain a proximal step.
- Replace this step by parametrized operators (not necessarily prox), where the parameters are learned.

Compressed Sensing using Generative Models

Generative Models:

- Examples are variational auto-encoder or generative adversarial networks (GANs)
- General: Neural networks

$$\mathbb{R}^k \ni z \mapsto G(z) \in \mathbb{R}^n,$$

where $k \ll n$ and G is trained to produce elements similar to training data.

Task:

- Let $A \in \mathbb{R}^{m \times n}$ Gaussian matrix, measurements $y = Ax_0 + \eta$ be given.
- Solve $z_0 \in \operatorname{argmin}_{z \in \mathbb{R}^k} \|AG(z) - y\|_2^2$ (non-convex) to within additive ε of optimum.

Theorem (Bora et al.; 2017): G generative model from d-layer ReLU neural network and $m \in \mathcal{O}(kd \log(n))$. Then with overwhelming probability

$$\|G(z_0) - x_0\|_2 \leq 6 \min_{z \in \mathbb{R}^k} \|G(z) - x_0\|_2 + 3\|\eta\|_2 + 2\varepsilon.$$

The “Best” Deep Learning Approach to Limited-Angle CT

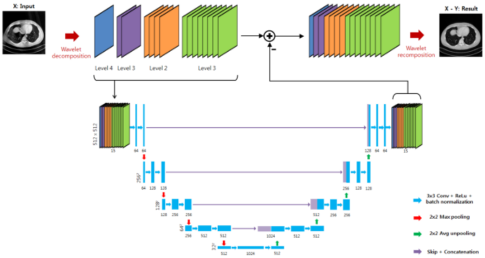


Image source: [Gu & Ye, 2017]:

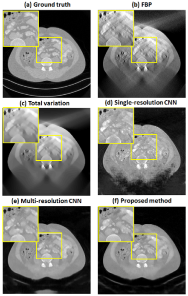


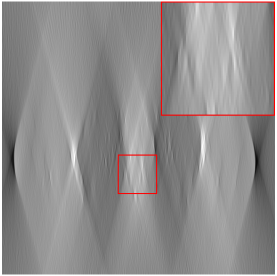
Image source: [Gu & Ye, 2017]:

- Missing theory, unclear what the neural network really does:
  - Entire image is processed!
  - Which features are modified?
  - Lack of a clear interpretation!
- The neural network needs to learn a lot of streaking artifacts (+noise)

[J. Gu and J. C. Ye. Multi-scale wavelet domain residual learning for limited-angle CT reconstruction. In: Procs Fully3D (2017), pp. 443447.]

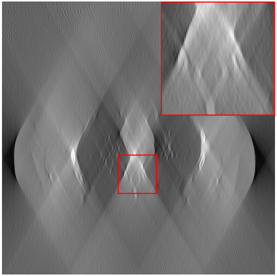
A True Hybrid Approach

Zooming in on the Recovery Problem



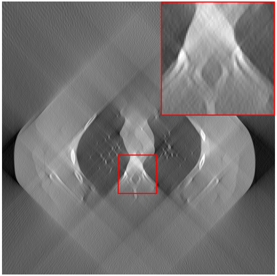
$\phi = 15^\circ$, filtered backprojection (FBP)

Zooming in on the Recovery Problem



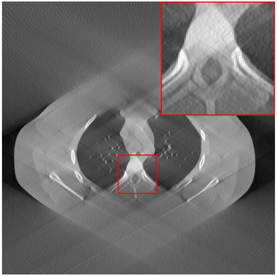
$\phi = 30^\circ$, filtered backprojection (FBP)

Zooming in on the Recovery Problem



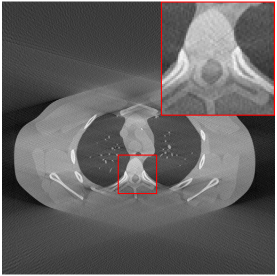
$\phi = 45^\circ$, filtered backprojection (FBP)

Zooming in on the Recovery Problem



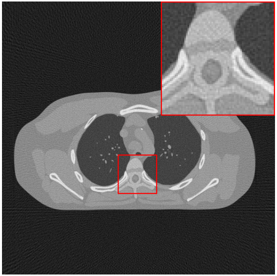
$\phi = 60^\circ$, filtered backprojection (FBP)

Zooming in on the Recovery Problem



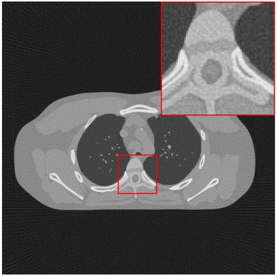
$\phi = 75^{\circ}$, filtered backprojection (FBP)

Zooming in on the Recovery Problem



$\phi = 90^\circ$, filtered backprojection (FBP)

Zooming in on the Recovery Problem



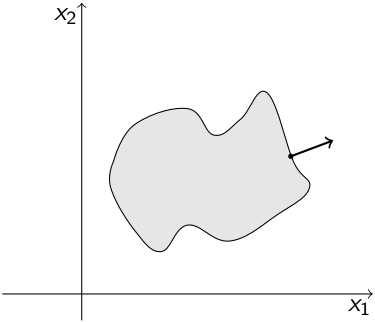
$\phi = 90^\circ$, filtered backprojection (FBP)

Some Observations:

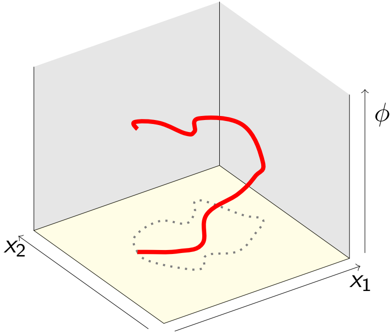
- Only certain boundaries/features seem to be *"visible"*!
- Missing wedge creates artifacts!
- Highly ill-posed inverse problem!

Fundamental Understanding of the Problem

This Phenomenon is well understood and mathematically analyzed via the concept of *microlocal analysis*, in particular, *wavefront sets*.



$f = \mathbb{I}_D$ for a set $D \subseteq \mathbb{R}^2$ with smooth boundary

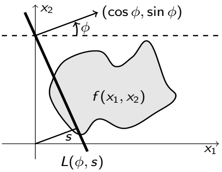


Visualization in phase space

Visibility in CT

Theorem ([Quinto, 1993]): Let $L_0 = L(\phi_0, s_0)$ be a line in the plane. Let $(x_0, \xi_0) \in \mathrm{WF}(f)$ such that $x_0 \in L_0$ and ξ_0 is a normal vector to L_0.

- The singularity of f at (x_0, ξ_0) causes a unique singularity in $W(\mathcal{R} f)$ at (ϕ_0, s_0).
- Singularities of f not tangent to $L(\phi_0, s_0)$ do not cause singularities in $\mathcal{R} f$ at (ϕ_0, s_0).

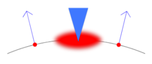


"visible": singularities tangent to sampled lines

"invisible": singularities not tangent to sampled lines

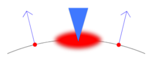
Shearlets can Help

Key Idea: Filling the missing angle is an inpainting problem of the wavefront set!

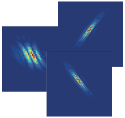


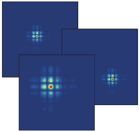
Shearlets can Help

Key Idea: Filling the missing angle is an inpainting problem of the wavefront set!



Theorem (K, Labate, 2006): "Shearlets can identify the wavefront set at fine scales."





More Precisely:

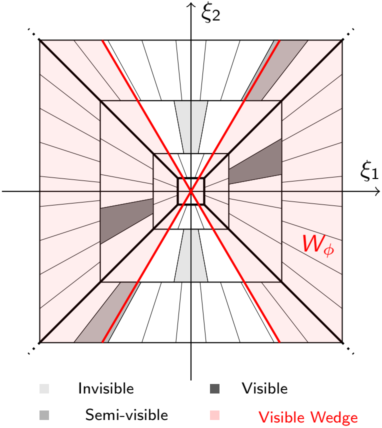
- Continuous Shearlet Transform:

$$L^2(\mathbb{R}^2) \ni f \mapsto \mathcal{SH}_\psi f(a,s,t) = \langle f, \psi_{a,s,t} \rangle, \quad (a,s,t) \in \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}^2.$$

- Resolution of Wavefront Sets (simplified from [K & Labate, 2006], [Grohs, 2011])

$$\mathrm{WF}(f)^c = \Big\{ (t_0, s_0) \in \mathbb{R}^2 \times [-1,1] : \text{for } (t,s) \text{ in neighborhood } U \text{ of } (t_0,s_0): \\ |\mathcal{SH}_\psi f(a,s,t)| = \mathcal{O}(a^k) \text{ as } a \longrightarrow 0, \forall k \in \mathbb{N}, \text{ unif. over } U \Big\}$$

Shearlets can Separate the Visible and Invisible Part



The High-level Idea

Avenue of Research

- Shearlets are proven to resolve the wavefront set.
- Use them in sparse/limited angle tomography for filling in missing parts of the wavefront set.

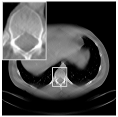
Practical Questions:

- How can we access the visible parts with shearlets?
  $\rightsquigarrow$ *Sparse Regularization!*
- How can we inpaint the missing parts?
  $\rightsquigarrow$ *Deep Learning!*

Our Approach "Learn the Invisible (LtI)"

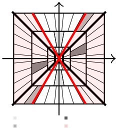
(Bubba, K, Lassas, März, Samek, Siltanen, Srinivan; 2018)

Step 1: *Reconstruct the visible*



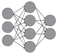
$$f^* := \operatorname{argmin}_{f \geq 0} \| \mathcal{R}_\phi f - g\|_2^2 + \| \mathrm{SH}_\psi(f)\|_{1,w}$$

- Best available classical solution (little artifacts, denoised)
- Access "wavefront set" via sparsity prior on shearlets:
  - For $(j,k,l) \in \mathcal{I}_{\mathtt{inv}}$: $\mathrm{SH}_\psi(f^*)_{(j,k,l)} \approx 0$
  - For $(j,k,l) \in \mathcal{I}_{\mathtt{vis}}$: $\mathrm{SH}_\psi(f^*)_{(j,k,l)}$ reliable and near perfect



Step 2: *Learn the invisible*

$$\mathcal{NN}_\theta : \ \mathrm{SH}_\psi(f^*)_{\mathcal{I}_{\mathtt{vis}}} \longrightarrow F \left(\overset{!}{\approx} \mathrm{SH}_\psi(f_{\mathtt{gt}})_{\mathcal{I}_{\mathtt{inv}}} \right)$$

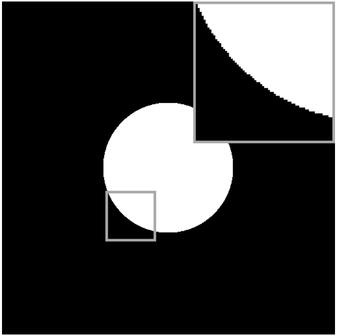


Step 3: *Combine*

$$f_{\mathtt{LtI}} = \mathrm{SH}_\psi^T (\mathrm{SH}_\psi(f^*)_{\mathcal{I}_{\mathtt{vis}}} + F)$$

Numerical Simulation

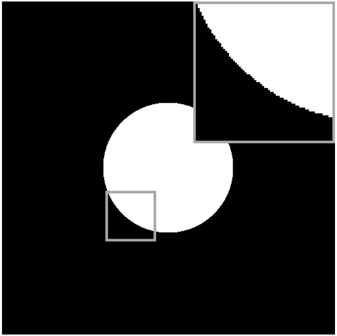
Verify the concept of (in-)visibility



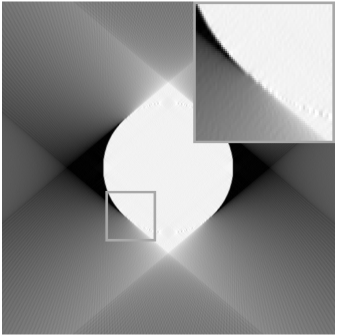
f_{gt}

Numerical Simulation

Verify the concept of (in-)visibility



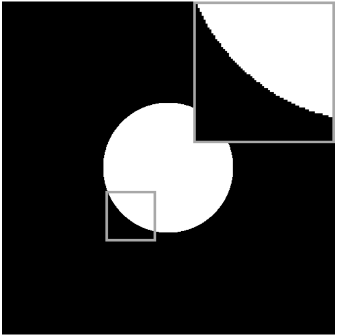
f_{gt}



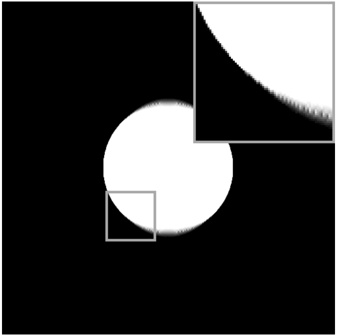
FBP

Numerical Simulation

Verify the concept of (in-)visibility

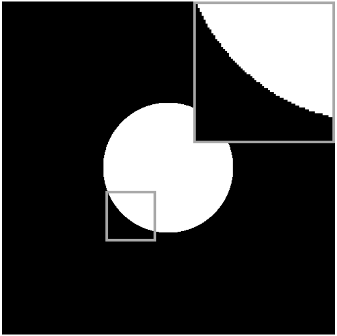


f_{gt}

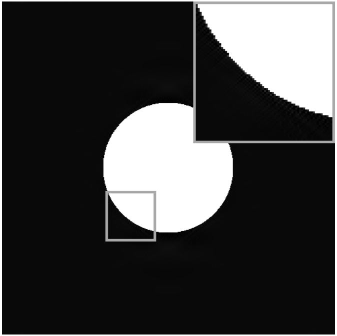


ℓ_1-analysis shearlet solution f^*

Verify the concept of (in-)visibility with the help of an oracle:



$f_{\mathtt{gt}}$

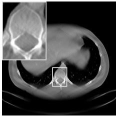


$\mathrm{SH}_\psi^T\left(\mathrm{SH}_\psi(f^*)_{\mathcal{I}_{\mathtt{vis}}} + \mathrm{SH}_\psi(f_{\mathtt{gt}})_{\mathcal{I}_{\mathtt{inv}}}\right)$

Our Approach "Learn the Invisible (LtI)"

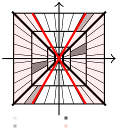
(Bubba, K, Lassas, März, Samek, Siltanen, Srinivan; 2018)

Step 1: *Reconstruct the visible*



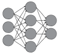
$$f^* := \operatorname{argmin}_{f \geq 0} \| \mathcal{R}_\phi f - g \|_2^2 + \| \mathrm{SH}_\psi(f) \|_{1,w}$$

- Best available classical solution (little artifacts, denoised)
- Access "wavefront set" via sparsity prior on shearlets:
  - For $(j,k,l) \in \mathcal{I}_{\mathtt{inv}}$: $\mathrm{SH}_\psi(f^*)_{(j,k,l)} \approx 0$
  - For $(j,k,l) \in \mathcal{I}_{\mathtt{vis}}$: $\mathrm{SH}_\psi(f^*)_{(j,k,l)}$ reliable and near perfect



Step 2: *Learn the invisible*

$$\mathcal{NN}_\theta : \ \mathrm{SH}_\psi(f^*)_{\mathcal{I}_{\mathtt{vis}}} \longrightarrow F \ \left(\overset{!}{\approx} \mathrm{SH}_\psi(f_{\mathrm{gt}})_{\mathcal{I}_{\mathtt{inv}}} \right)$$



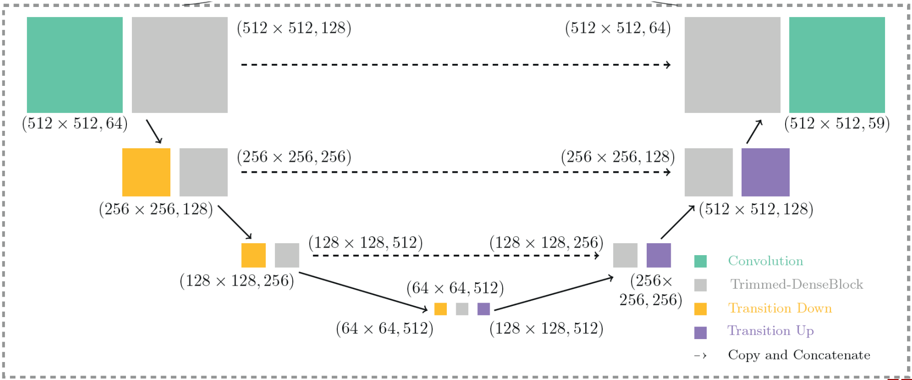
Step 3: *Combine*

$$f_{\mathtt{LtI}} = \mathrm{SH}_\psi^T \left(\mathrm{SH}_\psi(f^*)_{\mathcal{I}_{\mathtt{vis}}} + F \right)$$

Our Approach – Step 2: PhantomNet

U-Net-like CNN architecture $\mathcal{NN}_\theta$ (40 layers) that is trained by minimizing:

$$\min_\theta \frac{1}{N} \sum_{j=1}^{N} \|\mathcal{NN}_\theta(\mathrm{SH}(f_j^*)) - \mathrm{SH}(f_j^{\mathrm{gt}})_{\mathcal{I}_{\mathrm{inv}}}\|_{w,2}^2.$$



Learning the Invisible

Model Based & Data Driven: Only learn what needs to be learned!

Advantages over Pure Data Based Approach:

- Interpretation of what the CNN does ($\rightsquigarrow$ 3D inpainting)
- Reliability by learning only what is *not visible* in the data
- Better performance due to better input
- The neural network does not process entire image, leading to...
  - ...less blurring by U-net
  - ...fewer unwanted artifacts
- Better generalization

Disadvantage:

- Speed: dominated by ℓ^1-minimization

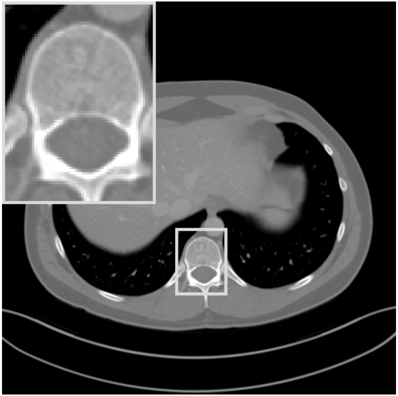
Setup

Experimental Scenarios:

- Mayo Clinic[1]: human abdomen scans provided by the Mayo Clinic for the AAPM Low-Dose CT Grand Challenge.
  - 10 patients (2378 slices of size 512×512 with thickness 3mm)
  - 9 patients for training (2134 slices) and 1 patient for testing (244 slices)
  - simulated noisy *fanbeam* measurements for 60° missing wedge
- Lotus Root: real data measured with the μCT in Helsinki
  - generalization test of our method (training is on Mayo data!)
  - 30° missing wedge
- ...

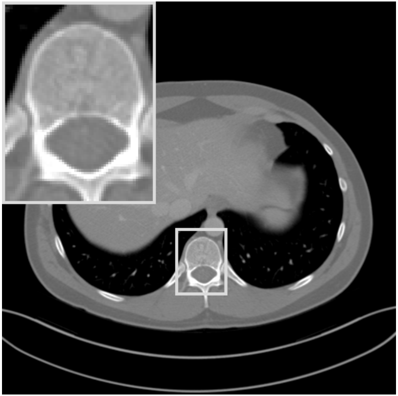
[1] We would like to thank Dr. Cynthia McCollough, the Mayo Clinic, the American Association of Physicists in Medicine (AAPM), and grant EB01705 and EB01785 from the National Institute of Biomedical Imaging and Bioengineering for providing the Low-Dose CT Grand Challenge data set.

Evaluation on Test Patient

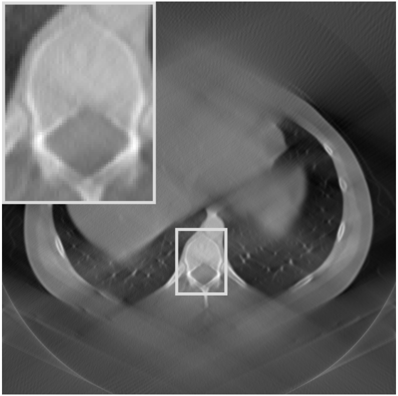


f_{gt}

Evaluation on Test Patient

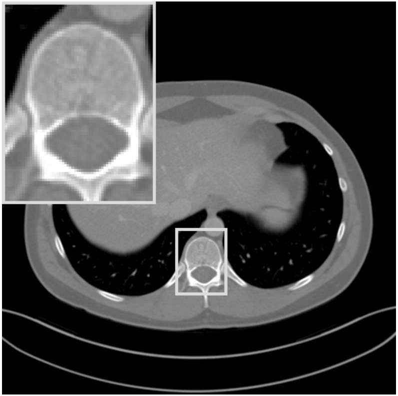


f_{gt}

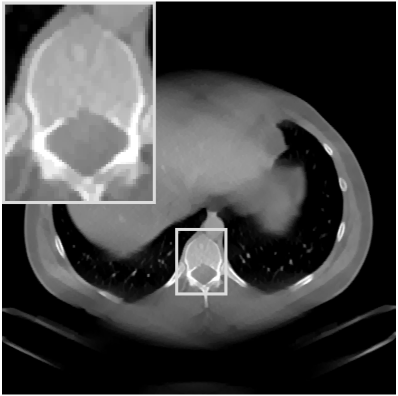


f_{FBP}: RE = 0.50, HaarPSI=0.35

Evaluation on Test Patient

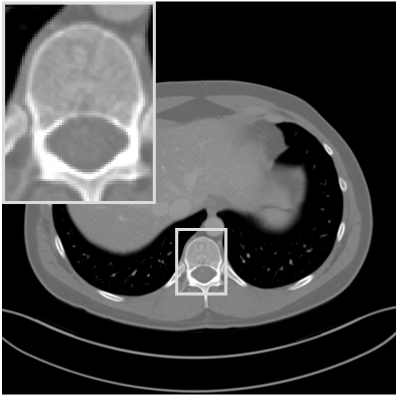


f_{gt}

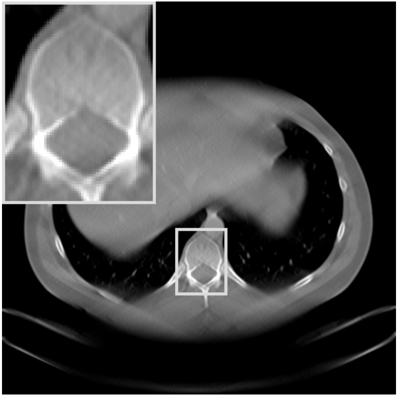


f_{TV}: RE $= 0.21$, HaarPSI$=0.41$

Evaluation on Test Patient

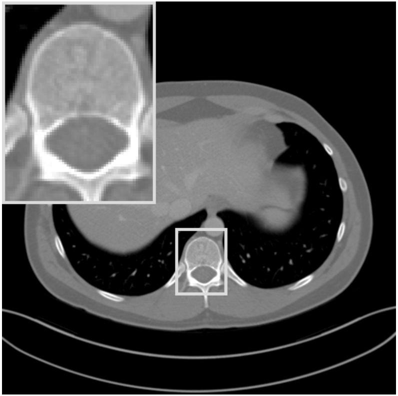


f_{gt}



f^*: RE $= 0.19$, HaarPSI$=0.43$

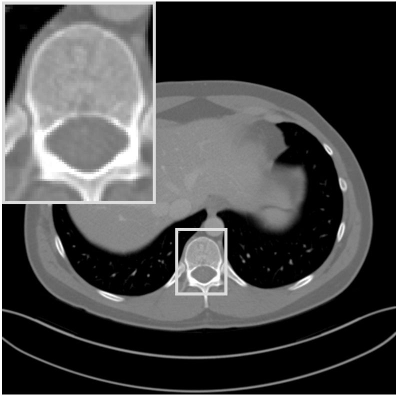
Evaluation on Test Patient



f_{gt}

$f_{\text{[Gu \& Ye, 2017]}}$: RE $= 0.22$, HaarPSI$=0.40$

Evaluation on Test Patient



$f_{\mathtt{gt}}$

$f_{\mathtt{LtI}}$: RE $= 0.09$, HaarPSI$=0.76$

Average over Test Patient

Method	RE	PSNR	SSIM	HaarPSI
f_{FBP}	0.47	17.16	0.40	0.32
f_{TV}	0.18	25.88	0.85	0.37
f^*	0.17	26.34	0.85	0.40
$f_{\text{[Gu \& Ye, 2017]}}$	0.25	23.06	0.61	0.34
$\mathcal{NN}_\theta(f_{\mathrm{FBP}})$	0.15	27.40	0.78	0.52
$\mathcal{NN}_\theta(\mathrm{SH}(f_{\mathrm{FBP}}))$	0.16	26.80	0.74	0.52
f_{LtI}	**0.08**	**32.77**	**0.93**	**0.73**

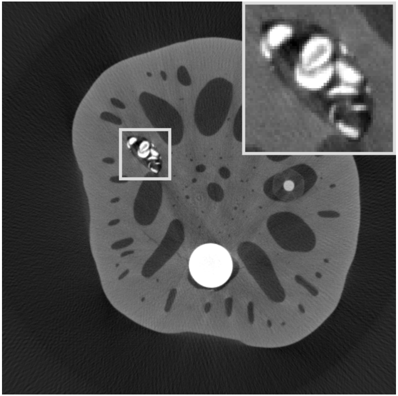
HaarPSI (Reisenhofer, Bosse, K, and Wiegand; 2018)

Advantages over (MS-)SSIM, FSIM, PSNR, GSM, VIF, etc.:

- Achieves higher correlations with human opinion scores.
- Can be computed very efficiently and significantly faster.

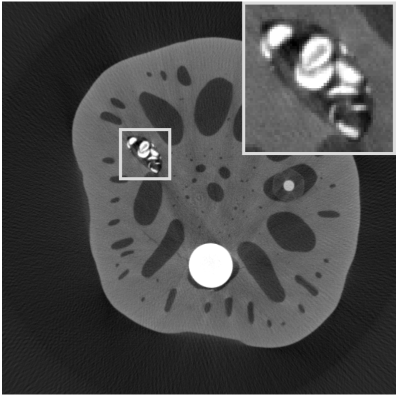
www.haarpsi.org

Generalization to Lotus Root

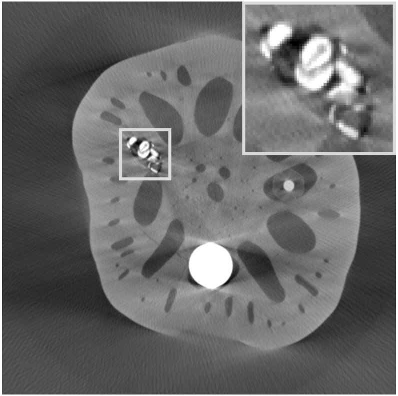


f_{gt}

Generalization to Lotus Root

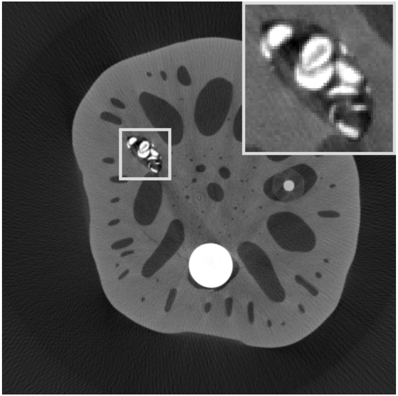


f_{gt}

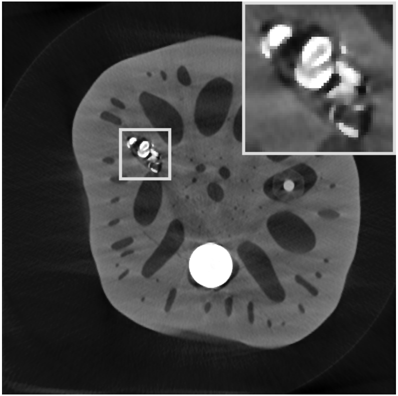


f_{FBP}: RE $= 0.31$, HaarPSI$=0.61$

Generalization to Lotus Root

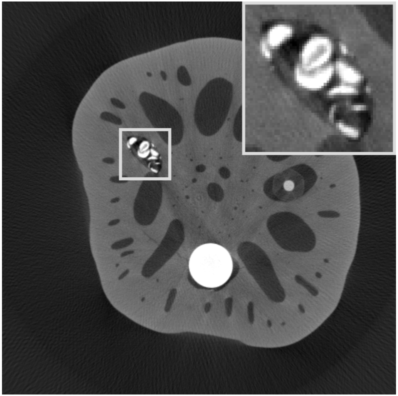


f_{gt}

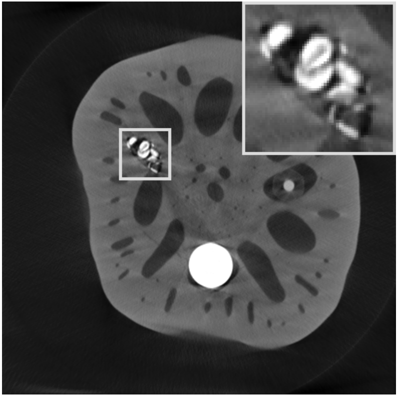


f_{TV}: RE $=$ 0.12, HaarPSI$=$0.74

Generalization to Lotus Root

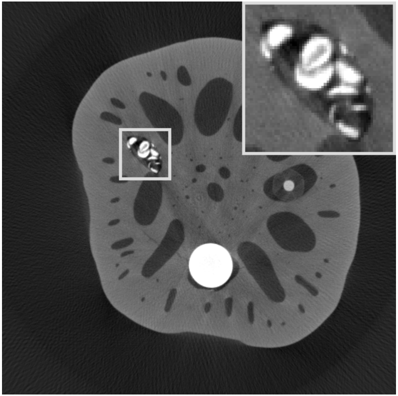


f_{gt}

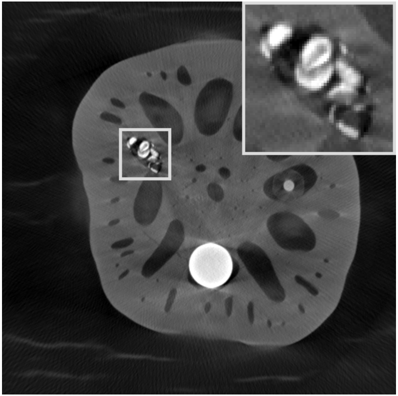


f^*: RE $= 0.11$, HaarPSI$=0.75$

Generalization to Lotus Root

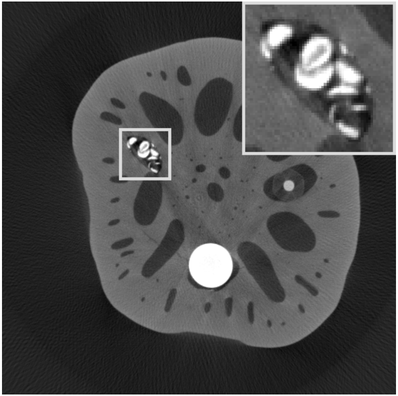


f_{gt}

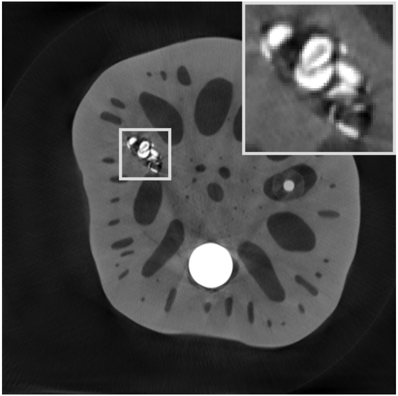


$f_{[\text{Gu \& Ye, 2017}]}$: RE $= 0.25$, HaarPSI$=0.62$

Generalization to Lotus Root



f_{gt}



$f_{\mathtt{LtI}}$: RE $= 0.11$, HaarPSI$=0.83$

Conclusions

What to take Home...?

Model-Based Side:

- *Inverse problems* can be solved by *sparse regularization*.
- *Shearlets* are optimal for imaging science problems.
- Methods based on *mathematical models* today often *reach a barrier*.

Deep Learning:

- Impressive performance *in combination with classical mathematical methods* (Inverse Problems, PDEs, ...).
- *Theoretical foundation* of neural networks *almost entirely missing*: *Expressivity, Learning, Generalization, and Interpretability*.

Combining Both Sides (Limited-Angle Tomography):

- Access and reconstruct the *visible part* using *shearlets*.
- Learn only the *invisible parts* with a *deep neural network*.

$\rightsquigarrow$ *Learning the Invisible (LtI)!*

THANK YOU!

References available at:

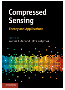
www.math.tu-berlin.de/∼kutyniok

Code available at:

www.ShearLab.org

Related Books:

- Y. Eldar and G. Kutyniok
  Compressed Sensing: Theory and Applications
  Cambridge University Press, 2012.



- G. Kutyniok and D. Labate
  Shearlets: Multiscale Analysis for Multivariate Data
  Birkhäuser-Springer, 2012.

- P. Grohs and G. Kutyniok
  Theory of Deep Learning
  Cambridge University Press (in preparation)