MIT Libraries | DSpace@MIT

MIT Open Access Articles

Noise-resilient deep learning for integrated circuit tomography

Massachusetts Institute of Technology

Check for updates

Noise-resilient deep learning for integrated circuit tomography

ZHEN GUO,[1],[†] 🄳 ZHIGUANG LIU,[2],[†] GEORGE BARBASTATHIS,[2,3,7]
QIHANG ZHANG,[1] MICHAEL E. GLINSKY,[4] BRADLEY K. ALPERT,[5]
AND ZACHARY H. LEVINE[6,8] 🄳

[1]*Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[2]*Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[3]*Singapore-MIT Alliance for Research and Technology (SMART) Centre, 138602, Singapore*
[4]*qiTech Consulting, Santa Fe, New Mexico 87505, USA*
[5]*Applied and Computational Mathematics Division, National Institute of Standards and Technology, Boulder, Colorado 80305, USA*
[6]*Quantum Measurement Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, USA*
[7]*gbarb@mit.edu*
[8]*zlevine@nist.gov*
[†]These authors contributed equally to this work.

Abstract: X-ray tomography is a non-destructive imaging technique that reveals the interior of an object from its projections at different angles. Under sparse-view and low-photon sampling, regularization priors are required to retrieve a high-fidelity reconstruction. Recently, deep learning has been used in X-ray tomography. The prior learned from training data replaces the general-purpose priors in iterative algorithms, achieving high-quality reconstructions with a neural network. Previous studies typically assume the noise statistics of test data are acquired *a priori* from training data, leaving the network susceptible to a change in the noise characteristics under practical imaging conditions. In this work, we propose a noise-resilient deep-reconstruction algorithm and apply it to integrated circuit tomography. By training the network with regularized reconstructions from a conventional algorithm, the learned prior shows strong noise resilience without the need for additional training with noisy examples, and allows us to obtain acceptable reconstructions with fewer photons in test data. The advantages of our framework may further enable low-photon tomographic imaging where long acquisition times limit the ability to acquire a large training set.

1. Introduction

X-ray tomography is a non-destructive imaging technique that visualizes the interior features of solid objects, with applications in biomedical imaging [1–3], materials science [4–6], manufacturing inspection [7,8], and other disciplines. The core idea is to measure the X-ray flux projected through the object at many angles, and then use a reconstruction algorithm to determine its attenuation coefficients throughout the volume. A high-quality reconstruction can be generated by densely-sampled, full-angle, and noise-free measurements that cover the entire Fourier space [9]. However, in practice, acquiring such measurements can be a challenge in situations where an object is not amenable to full angular inspection, such as with laminar structures, when the object is radiation sensitive, or when the acquisition is time sensitive. For measurements that are limited-angle, sparse, and noisy, the resulting inverse problem becomes ill-conditioned due to the deficits in the Fourier-space information [10–12]. Under such sampling

conditions, direct and filtered back-projection algorithms fail to produce satisfactory results. To retrieve a high-fidelity reconstruction, therefore, it is essential to utilize regularization [13], here provided by prior distributions.

Conventionally, the regularization prior is a general-purpose penalty function included in an iterative optimization algorithm. The algorithm starts from an initial reconstruction, and then progresses by iteratively updating the reconstruction to maximize its likelihood given the noisy tomographic measurements. This likelihood function is based on X-ray propagation, and is also called the data-fidelity term in the computed tomography (CT) community. The maximum likelihood estimate (MLE) assumes a prior by using a particular penalty function to the data-fidelity term, often to be a mean square error under a Gaussian noise prior on the measurements. In addition, the optimization may include another prior term that penalizes the lack of regularity on the reconstruction, leading to a maximum *a posteriori* (MAP) estimate. One of the most successful priors for tomography is sparsity, which assumes the reconstruction has a sparse representation after certain transformation [14–16]. Total Variation (TV), in particular, penalizes the discrete gradient of the reconstruction to minimize spatial variability [17,18].

Alternatively, the regularization priors can also be learned, e.g., through unsupervised machine learning such as sparse coding [19,20], or supervised machine learning methods [21–23] on well-labeled data. Given a paired dataset, a supervised machine learning algorithm can be trained to create an inverse map that directly generates object reconstructions from the tomographic measurements. Deep learning, a subset of machine learning based on artificial neural networks, has been particularly successful in approximating the inverse map. During training, high-order spatial correlations in the objects, as well as the noise statistics of the measurements, are implicitly learned. Test data, i.e., data which are disjoint from training data, are used to evaluate the generalization performance of the learned priors [24,25]. A recent trend in research is to also compute an estimate of the object by a conventional algorithm to prepare the training data, empirically achieving higher reconstruction quality by separating the forward model of tomography from training [26–28]. Among these approaches, filtered back-projection (FBP) with a convolutional neural network based on UNet is favored due to its computational efficiency. Once trained, networks with learned priors are able to greatly improve the FBP reconstruction without requiring additional iterations or regularization [29].

One major assumption in previous studies of the learned prior is that the noise statistics in the measurements are comparable between training and test datasets [28,30,31]. Assuming measurements are corrupted by Poisson noise, the reconstruction quality from using the learned prior is often evaluated by test data with the same number of photons per ray as the training data. However, in practical tomography systems, training and test data might have different photon counts per ray due to the variability in the light source and detector. For a test dataset with a different noise level than training data, generally referred to as out-of-distribution (OOD) data, the generalization of the learned prior is not guaranteed [32,33], leading to degradation in reconstruction fidelity [34–37]. The ability to maintain reconstruction quality against noise statistics variations in test data is called the "noise resilience" of the learned prior.

In this paper, we propose a noise-resilient deep-reconstruction algorithm for X-ray tomography and apply it to simulated and model integrated circuits. Our approach improves the noise resilience of the learned prior by using noise-resilient MAP reconstructions as the input to the neural network. Unlike previous efforts, we focus on the generalization of the deep learning algorithms to test data with different noise levels than the training data, which is critical in practical applications. By incorporating a Gaussian noise prior and a sparsity-promoting prior to the MAP reconstruction, our method reduces the distribution shifts between input reconstructions obtained from measurements with different noise levels. This leads to noise resilience in the learned prior without the need to sample the noise statistics using additional measurements, which is particularly useful in applications where acquiring training datasets at different noise

levels is challenging, such as in circuit imaging. Both simulation and experimental results show that a Total Variation regularizer on the maximum *a posteriori* estimate, instead of a more standard maximum likelihood or filtered back-projection approach, improves the noise resilience of the learned prior. Without training samples from different photon statistics, our MAP+UNet approach can produce acceptable reconstruction down to 80 photons per ray in simulations, and 214 photons per ray in experiments, allowing us to obtain comparable fidelity reconstructions with 8× fewer photons in simulation and 2.5× fewer photons in experiments than using the FBP+UNet approach for imaging integrated circuits.

The rest of the paper is organized as follows: First, we introduce the background of imaging integrated circuits, the forward model for tomography, and three reconstruction algorithms in Section 2. Next, we explain our noise-resilient approach based on neural networks in Section 3. Details of our evaluation methods are in Section 4. Simulation and experimental results are in Sections 5.1 and 5.2, respectively. Limitations and future works are discussed in Section 6. Concluding remarks are in Section 7. All acronyms for the algorithms investigated in this paper are defined in Table 1.

Table 1. Acronyms for the reconstruction algorithms.

Abbreviation	Definition
FBP	filtered back projection
MLE	maximum likelihood estimate
MAP/TV	maximum *a posteriori* estimate with Total Variation
FBP+UNet	improved FBP, MLE, or MAP/TV reconstruction
MLE+UNet	using the learned prior from UNet, shown in Fig. 2
MAP+UNet	

2. Background

2.1. Imaging integrated circuits

Integrated circuits are fundamental to the operation of all industrialized countries. In the United States, imaging these structures has been identified as a national goal to ensure the high-quality manufacturing of integrated circuits used in advanced computing and communication technologies [38]. Over the past two decades, advances have been made from initial tomographic reconstructions of integrated circuit test structures [39] to ptychographic imaging for larger circuit areas [40]. Despite this progress, photon-limited measurements will continue to be a challenge in imaging large circuits. For example, recently a laboratory-based experiment performed a reconstruction of a hundred square micrometers of integrated circuit interconnect using fewer than 100 photons per voxel [41]. Further reduction in the photon requirements will allow for the measurement of larger areas of interconnect irrespective of hardware limitations. Such considerations motivate this work.

2.2. Forward model for X-ray tomography

An X-ray tomography system commonly consists of a cone-beam source, a high-precision rotation stage, and a detector. The conceptual diagram is in Fig. 1. The measurements are projections of the attenuation coefficients of the object at different angles, i.e., integrals over rays from the source through the object to the detector pixels at each angle, and over different photon energies of each set of rays. For monochromatic illumination with reconstruction to a single material, the
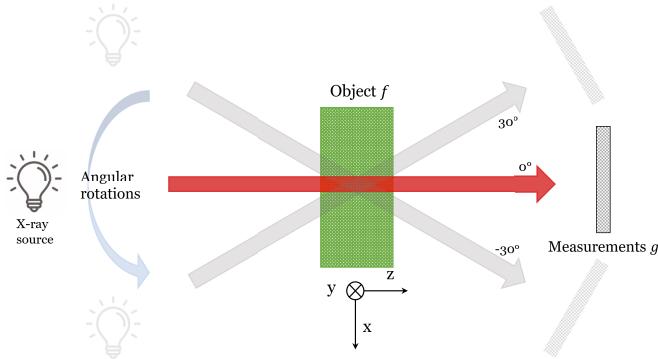
forward model can be simplified and discretized to

$$g = \mathbf{H}(f) = N_0 e^{-\alpha A f}, \tag{1}$$

where g is the set of noise-free measurements in terms of photon counts, $\mathbf{H}$ is the forward operator, N_0 is the photon count per ray from the source, α is the attenuation coefficient of a reference material with units of inverse length, and A is the system matrix with a unit of length, and f is the fractional density at a point in the object compared to the reference material. Here, α is a real scalar, Af is a matrix-vector product, and the exponential is applied to each component of the resulting vector. Therefore, the dimension of f is the number of voxels in the sample, the dimension of g is the number of X-ray paths through the sample—for all projections—that yield measurements, and an element of A specifies the length of a particular X-ray path through a particular voxel. In practice, measurements are subject to noise from various sources. Assuming Poisson statistics in the detection system, the forward model is modified as

$$g^* \sim \mathscr{P}(g), \tag{2}$$

where g^* is the noisy measurements vector, and $\mathscr{P}$ is a vector of Poisson distributions whose parameters are elements of the vector g. The operator $\sim$ means "is drawn from the distribution," following a convention from statistics. We implemented the forward model in PyTorch with parallel computation on a graphics processing unit [42].

Fig. 1. Conceptual diagram for the X-ray imaging system, shown in object-fixed coordinates, as is typical in medical tomography. The source and detector simultaneously rotate around the object along the y axis. X-rays, represented by red arrows, travel towards the pixelated detector shown as a dark grey rectangle. The X-rays, source, and detector are depicted in light grey to represent their position before and after the current rotation angle. In our experiment, however, the sample rotates with the source and detector in fixed positions.

2.3. Reconstruction algorithms

2.3.1. Filtered back-projection

The filtered back-projection (FBP) algorithm directly generates the inverse solution $\hat{f}$ by

$$\hat{f} = \mathbf{H}^{-1}\left[\mathbf{F}(g^*)\right]. \tag{3}$$

Here, $\mathbf{H}^{-1}$ is the back-projection operator and $\mathbf{F}$ is a filter to smooth the noisy measurements g^*, often chosen as a ramp at low frequencies and a tunable decay shape at higher frequencies up to the passband [43]. The FBP method is computationally efficient and effective for full-angle and densely sampled data [44]. However, reconstruction artifacts might appear for measurements that are limited in their total angular range, or sparsely sampled and noisy [45,46].

2.3.2.   Maximum likelihood and maximum *a posteriori* estimates

A maximum likelihood estimate (MLE) is the reconstruction obtained by iteratively maximizing an objective function based on a likelihood incorporating the projective geometry and a prior on the noise statistics of the measurements. We find an optimal estimate $\hat{f}$ using the objective function

$$\hat{f} = \arg\max_{f^{(0)}} \left[p(g^*|f^{(0)}) \right], \tag{4}$$

where $p(g^*|f^{(0)})$ is the likelihood of a proposed reconstruction $f^{(0)}$ given the noisy measurements g^*, representing the data-fidelity term. The estimate of the likelihood includes a prior by using a particular penalty function, often taken to be a mean square error which assumes additive Gaussian noise in the measurements.

When the iterative optimization includes an additional term that penalizes lack of regularity or mismatch to a prior distribution on the reconstruction, the objective function becomes [47,48]

$$\hat{f} = \arg\max_{f^{(0)}} \left[p(g^*|f^{(0)}) + \beta\Psi(f^{(0)}) \right], \tag{5}$$

where Ψ is the regularization prior on the proposed reconstruction, and β is the regularization parameter. The final reconstruction given the prior distribution on the reconstruction is the maximum *a posteriori* (MAP) estimate. The key to a high-quality iterative reconstruction is a proper choice of Ψ and β for a given set of objects [49–51]. One of the most successful priors for tomography is sparsity, which assumes the reconstruction has a sparse representation after certain transformation [14–16]. For a sparsity-promoting prior such as Total Variation (TV), the Ψ term penalizes variation in the recovered function, which tends to impose sparsity in the function's gradient. When the prior distribution is uniform or β is zero, the maximum *a posteriori* estimate is equivalent to the maximum likelihood estimate. In general, iterative optimization achieves higher reconstruction quality than direct inversion algorithms at the expense of increased computational cost [50,52]. While the regularization prior and its parameters can sometimes be chosen by objective criteria [53,54], in practice, they are usually determined by trial and error.

2.3.3.   Deep-reconstruction network from supervised learning

Supervised learning is a data-driven approach. Given a set of training samples from the joint distribution $P(g^*, f_0)$, where g^* is drawn from the set of physical measurements, and f_0 is drawn from the set of ground truth objects, the machine learning algorithm acquires the regularization prior implicitly by approximating $P(f_0 \mid g^*)$, the conditional probability of having an output f_0 given an input g^*. In particular, deep learning, a subset of machine learning based on artificial neural networks, has shown promising results for tomography in recent years [23,26,55–58]. The process of approximating $P(f_0 \mid g^*)$ becomes an optimization problem, where the objective is to find a deep neural network with appropriate weights which can map the physical measurements g^* to its corresponding ground truth object f_0 from training data

$$\hat{\mathbf{w}}_g = \underset{\mathbf{w}}{\arg\min}\, \mathbb{E}_{(g^*, f_0)} \left[\mathscr{L}\{G_{\mathbf{w}}(g^*), f_0\} \right]$$
$$\approx \underset{\mathbf{w}}{\arg\min} \sum_{i=1}^{n_{\text{train}}} \mathscr{L}\left\{ G_{\mathbf{w}}(g_i^*), f_{0i} \right\}. \tag{6}$$

Here, $G_{\mathbf{w}}$ is a deep neural network, $\mathbf{w}$ is a vector of network parameters, i.e., weights, n_{train} is the number of training samples, and $\mathscr{L}$ is the objective function. This is known as end-to-end training [59]. Once trained, the inversion of the forward operator, the noise statistics of the measurements, and the object prior are, in principle, incorporated into the network parameters.

In contrast to end-to-end training, in physics-assisted training [28,29,60–62], a conventional reconstruction algorithm pre-processes the measurements in the training data before sending them into the neural network. Utilizing approximate reconstructions from a conventional algorithm reduces the learning burden, and leads to

$$\hat{\mathbf{w}}_f = \underset{\mathbf{w}}{\arg\min} \, \mathbb{E}_{(\hat{f}, f_0)} \Big[\mathcal{L}\{G_{\mathbf{w}}(\hat{f}), f_0\} \Big]$$

$$\approx \underset{\mathbf{w}}{\arg\min} \sum_{i=1}^{n_{\text{train}}} \mathcal{L}\left\{ G_{\mathbf{w}}(\hat{f_i}), f_{0i} \right\},$$

(7)

where $\hat{f}$ is an estimate of the object from a conventional algorithm.

One of the most popular physics-assisted methods is FBP with a convolutional neural network based on UNet, where the approximate reconstruction $\hat{f}$ is from the FBP algorithm to encapsulate the physical model of the imaging system. After training, the convolutional UNet can remove artifacts while preserving image structure in the FBP reconstruction without requiring additional iterations and regularization, showing high computational efficiency and low latency [29]. Other variants of this technique have also shown promising results with short inference time and reconstruction quality superior to MAP with general-purpose priors [22,26,63,64].

3. Noise resilience of deep-reconstruction networks

One major concern with using supervised learning for tomography is the generalization problem, which refers to the ability of the learned prior to adapt to unseen test data. When properly trained, the learned prior can generalize well and produce high-quality reconstructions for test data sampled from the same distribution as the training data. However, in practical tomography systems, training and test data can come from different distributions. Though one could know the class of imaging objects *a priori* and fix the imaging geometry, the noise characteristics in the data might vary with time.

Previous studies of supervised learning generally assume that the noise statistics in the measurements are comparable between training and test datasets. If measurements g^* are corrupted by Poisson noise, the reconstruction quality from the learned prior is evaluated by test data with the same photon count N_0 per ray as the training data. When the trained network is given test data collected at N_1 photons per ray, a noise distribution shift would occur between training and testing, and the reconstruction quality from the learned prior might not be guaranteed when no training samples at N_1 photons are provided. In other words, the estimated conditional probability $P(f_0 \,|\, g^*)$ may be unreliable if the joint distributions $P\left(g_{N_0}^*, f_0\right) \not\approx P\left(g_{N_1}^*, f_0\right)$, where N_0 and N_1 indicate the different photon statistics in the measurements g^*. The ability to maintain reconstruction quality against noise statistics variations in test data is regarded as the noise resilience of the deep-reconstruction network.

To improve the noise resilience of the deep reconstruction network, one could approximate $P(f_0 \,|\, g^*)$ by using a set of distributions $P\left(g_{N_i}^*, f_0\right), N_i \in \{0, 1, 2, \ldots\}$, resulting in a network that generalizes to the noise statistics. Also, one could approximate a series of $P\left(f_0 \,|\, g_{N_i}^*\right), N_i \in \{0, 1, 2, \ldots\}$, resulting in a bank of networks specialized in each noise level. However, these approaches require collecting training data under various noise levels, which is time-consuming for tomography systems with long acquisition times. Instead, our approach is to train the network using a joint distribution without collecting more noisy data. By utilizing a Gaussian prior on the noise statistics and a sparsity-promoting prior on the reconstruction, we produce noise-resilient $\hat{f}_{\text{MAP}}$ from the maximum *a posteriori* estimate, and then approximate $P\left(f_0 \,|\, \hat{f}_{\text{MAP}}\right)$ by sampling

the training distribution $P\left(\hat{f}_{\text{MAP}}, f_0\right)$ only at a given photon statistics. Test data from different photon statistics are used to evaluate the noise resilience of the learned prior.

4. Evaluation methods

4.1. Description of X-ray tomography experiment

Our X-ray imaging system is the Zeiss Xradia 620 Versa at MIT.nano. The X-ray spectrum is generated from a tungsten target with a tube voltage of 80.0 kV and a power of 10.0 W. The 3D printed sample from CircuitFaker [28], a generator of 3D spatial patterns that match rudimentary circuit-like statistics, is placed 230 mm from the source and is mechanically rotated from 0° to 360° evenly with 1600 projections by a high-precision stage. A charge-coupled device (CCD) is placed 572 mm from the source. The Fresnel number is approximately 100. Therefore, it is safe to use projections as the forward model. The imaging geometry is shared between simulations and experiments. With a cone angle of 3° (or maximum divergent angle 1.5°), and a maximum tolerance angle around 2.4° for the sample, we also conclude that it is safe to use the parallel-beam (Radon) assumption in the forward model. The Poisson statistics for the measurements in the simulations are synthetically generated. The experimental exposure time varies from 35 ms to 9 s per view to implement increasing photon counts.

4.2. Reconstruction algorithms for comparison

There are three traditional reconstruction algorithms, namely FBP, MLE, and MAP/TV, and three learning-based algorithms by adding a UNet for each. MAP+UNet is our noise-resilient approach by approximating $P\left(f_0 \,|\, \hat{f}_{\text{MAP}}\right)$, whereas MLE+UNet and FBP+UNet represent alternative learning-based algorithms by approximating $P\left(f_0 \,|\, \hat{f}_{\text{MLE}}\right)$ and $P\left(f_0 \,|\, \hat{f}_{\text{FBP}}\right)$, respectively. All acronyms for the algorithms are defined in Table 1. For a fair comparison, the three learning-based algorithms share the same optimization parameters and the UNet architecture. The only difference is in the input reconstruction. The regularization parameter β is 2 for all the MAP reconstructions at different noise levels. A conceptual diagram is shown in Fig. 2. For simulation, the weights in the networks are randomly initialized and then updated by 10 000 noise-free training samples. For the experiment, the weights are initialized by values trained from simulated data and then updated by 120 experimental training samples. This training strategy, called transfer learning, is useful to reduce the required number of experimental training samples [65].
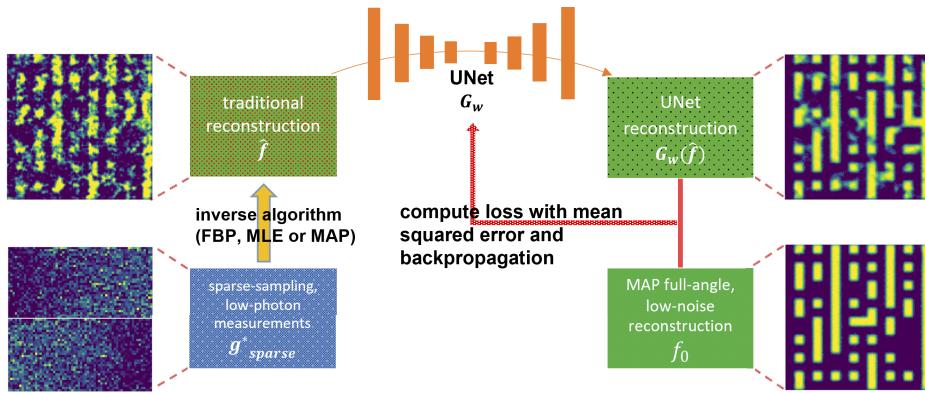
4.3. Network architecture

Figure 3 shows the network architecture for the learning-based algorithms. Four downsampling blocks and four up-sampling blocks were used in the UNet-like architecture. The spatial dimension of the feature map is reduced or up-sampled by 2 per block. The downsampling is achieved by stride convolution and the up-sampling is by transposed convolution. The initial input is $128 \times 128 \times 1$, the same as the output dimension. The latent vector in the center has the dimension $8 \times 8 \times 512$, where the channel size is in the last dimension. ReLU is used as the activation function in the blocks. The skip connections concatenate features from downsampling blocks to up-sampling blocks. The script that generates the architecture can be found on our GitHub page [42].
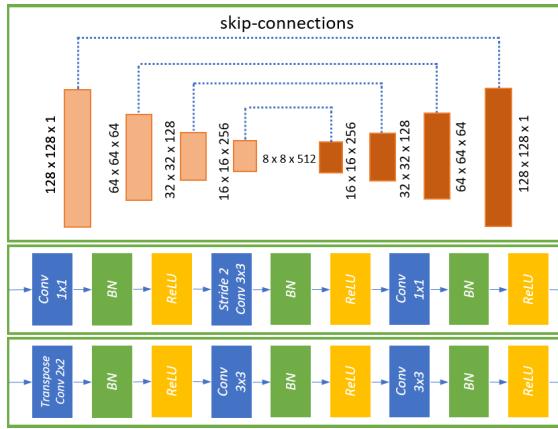
4.4. Training and test data preparation

4.4.1. Simulated data

Our simulated imaging objects are samples from CircuitFaker. The ground truth reconstructions are obtained from full-angle, densely-sampled (with 1600 angular views), and noise-free

Fig. 2. A conceptual diagram for the learning-based algorithms. An inverse algorithm first produces the traditional reconstruction $\hat{f}$ from the sparsely-sampled and low photon measurements. Then the UNet takes $\hat{f}$ and outputs reconstruction $G_{\mathbf{w}}(\hat{f})$. The differences between $G_{\mathbf{w}}(\hat{f})$ and f_0 are backpropagated to find the optimal UNet parameters based on Eq. (7).

Fig. 3. UNet architecture for the learning-based algorithms. The top box shows the overall design of the network, where the light orange modules are downsampling blocks and the dark orange modules are up-sampling blocks. Blue dotted lines represent the skip connections. The middle box shows the design of the downsampling blocks. The bottom box shows the design of up-sampling blocks. BN is batch normalization.

measurements with MAP. The training datasets are generated at a fixed imaging geometry (32 evenly spaced angular views) with 10 000 CircuitFaker objects using noise-free measurements. The test datasets are generated at the same imaging geometry with 1000 objects with different Poisson noise levels in the measurements, over the range of 32 to 2000 photons per ray.

Under the parallel-beam approximation, the cone-beam projections are divided into line projections with dimensions of 256×1. The reconstructions have $150 \times 150 \times 1$ voxels filling a region of 8.0 mm $\times$ 8.0 mm $\times$ 0.1 mm volume from the line projections. Although we simulate data for 16 layers of circuit objects in a single simulation, we reconstruct these layers independently, using a 2D algorithm for each. The reconstructions are cropped to 128×128 pixels for visual and quantitative comparison.

4.4.2. Experimental data

The experimental imaging objects are designed in OpenSCAD, with 3D configuration based on new samples drawn from CircuitFaker. Each circuit independently occupies one layer and stands on a broad substrate, and substrates are excluded from the datasets. The 3D configuration is fabricated by using projection stereo-lithography apparatus (Ember 3D printer, Autodesk) with clear resin (PR48), followed by washing with isopropyl alcohol to remove uncured monomers from the as-printed sample, then transferred to the tomography imaging system for experimental measurements. The cone-beam projections are cropped to the region of interest, downsampled, and then divided into line projections with dimensions of 256×1. The dimensions of the circuit reconstructions are the same as in the simulation.

The ground truth reconstructions are obtained from full-angle and densely-sampled (with 1600 angular views) measurements with the highest photon count per ray (13 598) with MAP. The training dataset consists of 120 samples collected with the highest photon count per ray with 32 evenly spaced angular views across the full range of angles. The test data sets consist of 40 samples each, collected over the range of 66 to 3347 photons per ray with the same angular views.

4.5. Algorithmic details

Our algorithms are implemented in Python 3.8.13 using PyTorch 1.12.1, and performed on MIT Supercloud [66] with Intel Xeon Gold 6248 and an NVIDIA V100 GPU. The CircuitFaker parameters are consistent with [28]. For all the deep learning algorithms, the number of trainable parameters is around 14 million. To train the networks, an AdamW optimizer [67] is used with parameters $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight decay of 0.05. The training objective function is the mean squared error. For training with simulated data, the batch size is 64, the number of warmup epochs is 40, and the total number of epochs is 400. The neural network receives one update from one batch of training samples per iteration, and all updates from all the batches once per epoch. For transfer learning on experimental data, the batch size is 10, the warmup epoch is 10, and the total number of epochs is 100. The initial learning rate is 5×10^{-4} per iteration, and the scheduler reduces the learning rate with half-cycle cosine after warmup, in proportion to $1 + \cos(\pi n / n_{\text{tot}})$ where n is the epoch number [68].

For the iterative reconstruction algorithm, an Adam optimizer [69] is used with an initial learning rate of 10^{-2}. The total number of iterations is 100, and the scheduler reduces the learning rate by half per 20 iterations. The objective function is again the mean squared error of the measurements. Importantly, the regularization parameter β for TV in MAP reconstruction is 0.25 for ground truth reconstructions, and 2 for all the test reconstructions at different noise levels. The FBP algorithm is imported from MATLAB with a Hanning filter [70].

4.6. Quality metrics and their acceptability thresholds

4.6.1. Pearson correlation coefficient

The Pearson correlation coefficient r is defined as

$$r_{f_0, \hat{f}} = \frac{\text{cov}(f_0, \hat{f})}{\sigma_{f_0} \, \sigma_{\hat{f}}}, \tag{8}$$

between ground truth reconstruction f_0 and reconstruction $\hat{f}$ from a particular algorithm, where cov is the covariance and σ is the standard deviation. This first metric is introduced to evaluate the perceptual quality of the reconstruction, and the acceptable quality threshold is $1 - r \leq 10^{-1}$, indicating a strong linear relationship between the reconstructions [71]. Its shortcoming is that it is a pixel-by-pixel correlation that is sensitive to misregistration and image distortion, and is not highly sensitive to the connectivity or topology of the image [72].

4.6.2. Mallat Scattering Transformation

The normalized L^2 distance of the logarithm of the Mallat Scattering Transform (MST) space [73,74] is defined as

$$\varphi_{f_0, \hat{f}} = \frac{\|\Phi(f_0) - \Phi(\hat{f})\|^2}{\|\Phi(f_0)\| \, \|\Phi(\hat{f})\|}, \tag{9}$$

where Φ is the logarithm of the MST operator. This second metric is introduced to evaluate multi-scale correlations and topology (connectivity) of the reconstructions. It is also insensitive to misregistration and image distortion. The acceptable quality threshold is $\varphi \leq n/M^2 = 3 \times 10^{-3}$, where the reconstruction is of size $M^2 = 128 \times 128$ and $n = 50$ is the average number of circuit elements.

MST can be viewed as a Convolutional Neural Network (CNN) with predetermined weights. The filters are designed so that the CNN can span an exponentially large range in scale with a kernel of constant size. Following [75], we define the logarithm of MST of an input image f as

$$\Phi(f) = \left(\log \Phi_J^{(0)}(f), \log \Phi_J^{(1)}(f), \log \Phi_J^{(2)}(f) \right) \tag{10}$$

and

$$\begin{aligned}
\Phi_J^{(0)}(f) &= \phi_J \circledast f \\
\Phi_J^{(1)}(f) &= \phi_J \circledast \left| \psi_{\lambda_1} \circledast f \right|, & \lambda_1 \in \Lambda_1, \\
\Phi_J^{(2)}(f) &= \phi_J \circledast \left| \psi_{\lambda_2} \circledast \left| \psi_{\lambda_1} \circledast f \right| \right|, & \lambda_1 \in \Lambda_1, \lambda_2 \in \Lambda_2.
\end{aligned} \tag{11}$$

Here, $\circledast$ denotes the convolution in 2D space, ϕ_J is a low-pass filter, $\{\psi_{\lambda_i}\}$ is a family of band-pass filters, $i = 1, 2$. Morlet filters [76] are used in our computation. There are three parameters that determine how the MST is taken: M^2 is the number of pixels in the image, J is the $\log_2$ of the scattering scale and L is the number of angles used in the transform. For this work, values of $M = 128$, $J = 4$, and $L = 8$ were used.
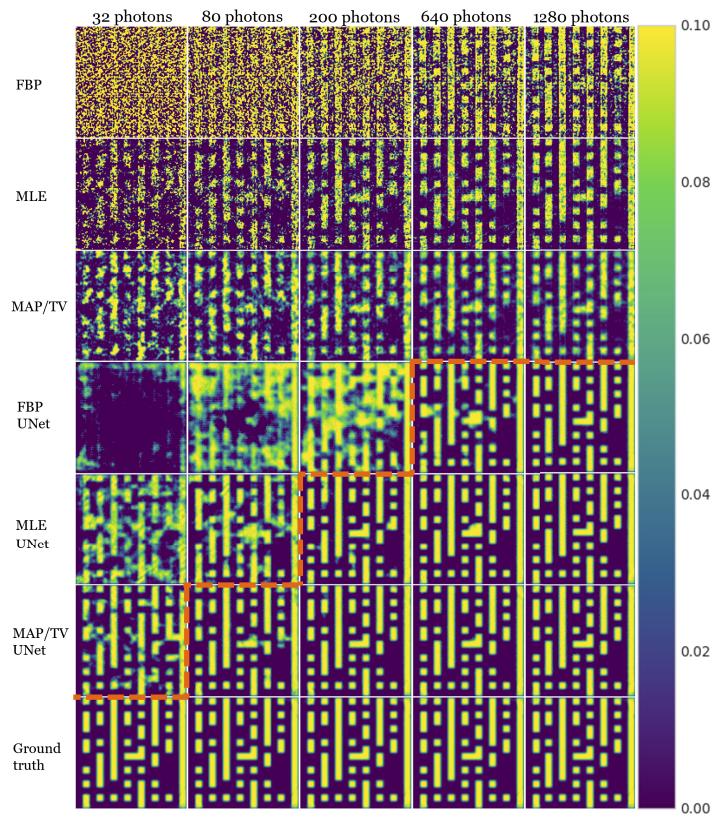
MST has been shown by Mallat [73] to be Lipschitz continuous to diffeomorphic deformation and invariant under translation. Therefore, MST is insensitive to misregistration and deformation because they are both small diffeomorphic deformations. It also induces a high sensitivity to topology, since the topology is invariant to diffeomorphic deformation. Furthermore, taking the logarithm of the transformation flattens the extracted features to a low-dimensional complex linear subspace where the topology of the features is exposed. The practical outcome is that the extracted features in MST space will form a high-precision cluster. The precision should be the fractional dimension of the space, n/M^2, where n is the number of the features.
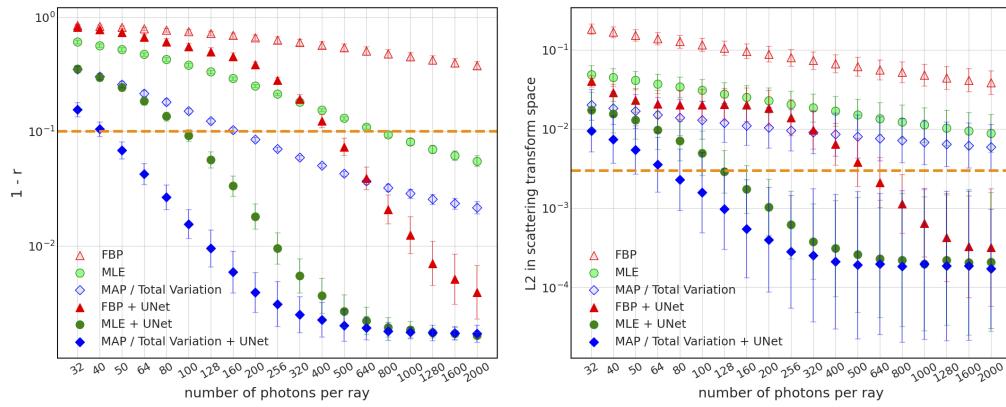
5. Results

5.1. Simulations

We first demonstrate our noise-resilient approach for ill-conditioned tomography with simulation. The imaging condition is in full-angle, sparse sampling (with 32 out of 1600 angular views, evenly spaced), and low-photon tomographic simulations where ill-conditioning exhibits itself severely without regularization. Figure 4 shows 2D reconstructions, at selected photon counts per ray, for different algorithms using simulated data. Each row represents different reconstruction algorithms. Each column represents reconstructions under different photon counts per ray. Visually, MAP+UNet generates better reconstructions at lower photons per ray compared to MLE+UNet and FBP+UNet approaches.
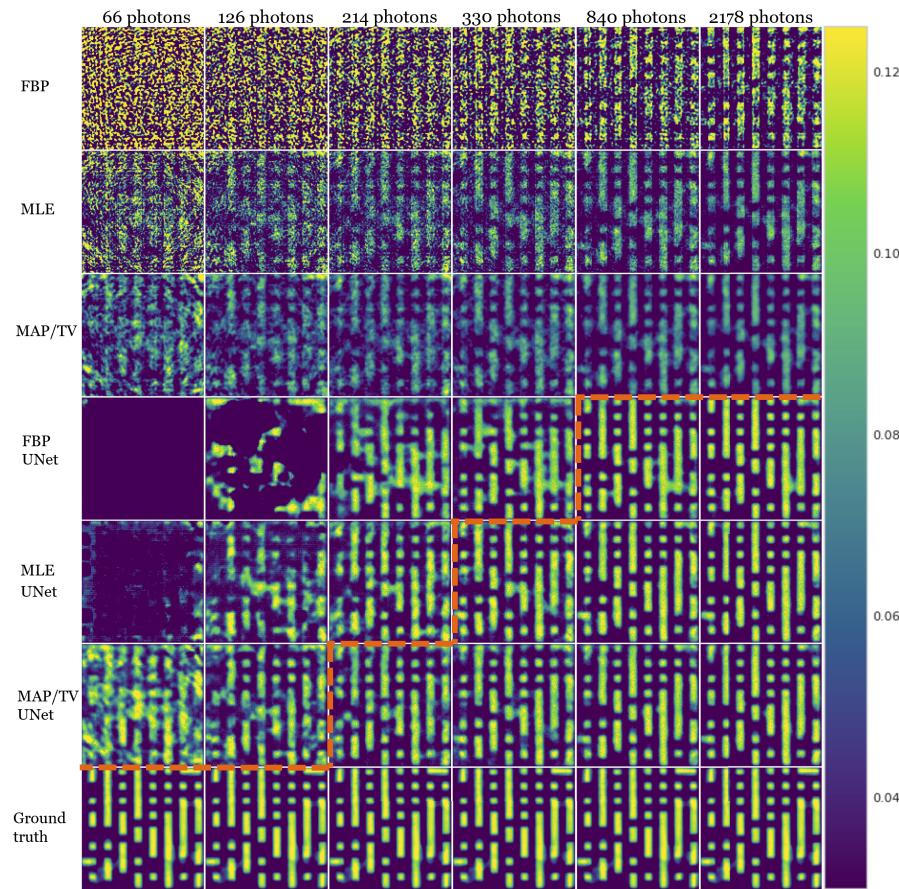
The quantitative comparison of different algorithms is shown in Fig. 5 where we report the means and standard errors for the two metrics over 1000 test instances from CircuitFaker. The general trends from the two metrics are similar, and we focus our interpretations below on the MST metric. Given the threshold of 3×10^{-3}, our MAP+UNet method satisfies the requirement

Fig. 4. Selected 2D reconstructions (in 128×128) for different algorithms using simulated data. Each row represents a reconstruction algorithm, and each column represents an intensity of the photon rays. The ground truth is repeated in the last row. The dotted orange line is the boundary between acceptable and unacceptable performance as determined by the MST metric.

Fig. 5. Quantitative comparison between different reconstruction algorithms for tomographic simulations under different photon counts per ray. The *x* axis is the number of photons per ray, and the *y* axis on the left figure is $1 - r$ where r is the Pearson correlation coefficient. The *y* axis on the right is the L^2 distance in MST. The error bars are standard deviations in the log scale of 1000 test instances. The dotted orange line shows the thresholds of acceptable performance.

Fig. 6. Selected 2D reconstructions (in 128 × 128) for different algorithms using experimental data. Each row represents a reconstruction algorithm. Each column represents an intensity of the photon rays. The dotted orange line is the boundary between acceptable and unacceptable performance as determined by the MST metric.
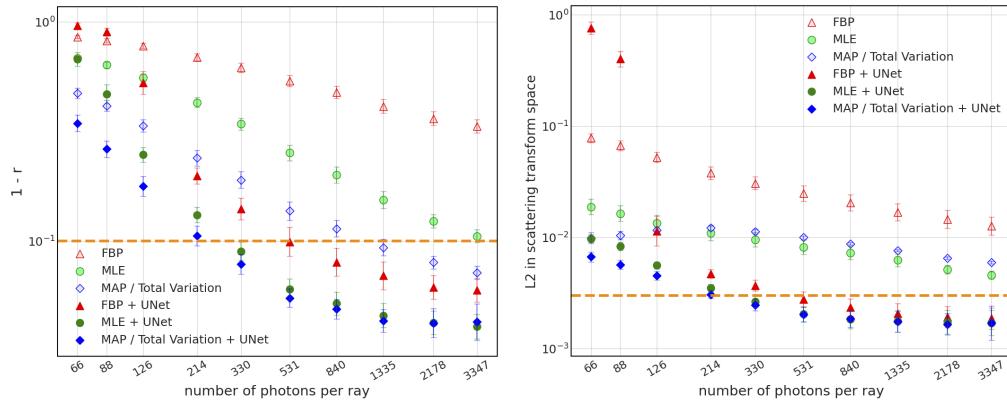
for fluxes larger than 80 photons per ray. The MLE+UNet method is the second-best algorithm, which can satisfy the requirement for 128 photons per ray or above. FBP+UNet exhibits the least noise resilience among the learning-based algorithms, withstanding noisy measurements only down to 640 photons per ray. Thus, MAP+UNet allows us to use 8× fewer photons than FBP+UNet in simulation. At low photon fluxes, the performance of the sparsity-promoting MAP+UNet is greater than that of the maximum-likelihood-based MLE+UNet. At high photon flux, the discrepancy between the two algorithms is eliminated. Finally, compared to the results of each traditional reconstruction algorithm, the addition of learning with a UNet leads to a lower photon flux requirement.

5.2. Experiments

Next, we demonstrate our noise-resilient approach to experimental data, using the same imaging condition as in the simulation. Figure 6 shows 2D reconstructions, at selected photon counts per ray, for different algorithms using experimental projection data.

Figure 7 shows the quantitative comparison between different algorithms for experimental data collected at various photon fluxes, based on the two quality metrics. The trends seen in the

simulation are apparent in the experimental results as well: a more faithful input reconstruction leads to an equal or better reconstruction from UNet, with equality more likely at higher photon counts. At lower photon counts, MAP+UNet is the most noise-resilient learning-based algorithm. Using the same requirement for acceptable quality as the simulation ($\varphi \leq 3 \times 10^{-3}$ for the MST metric), none of the traditional algorithms is acceptable for sparsely sampled and low-photon experimental measurements. In contrast, each of the learning-based algorithms yields an acceptable performance. The thresholds for acceptable performance for all the learning-based algorithms in the experiment and simulation are summarized in Table 2. Quantitative results using alternative metrics can be found in the Appendix (Fig. 8). Overall, MAP+UNet allows us to use 2.5× fewer photons than FBP+UNet in experimental data, vs. an 8× reduction in simulation.

Fig. 7. Quantitative comparison between different reconstruction algorithms for experimental data of 40 instances under different photon counts per ray. Symbols and error bars as in Fig. 5. The dotted orange lines show the thresholds of acceptable performance.

Table 2. Thresholds for acceptable performance based on the MST metric.

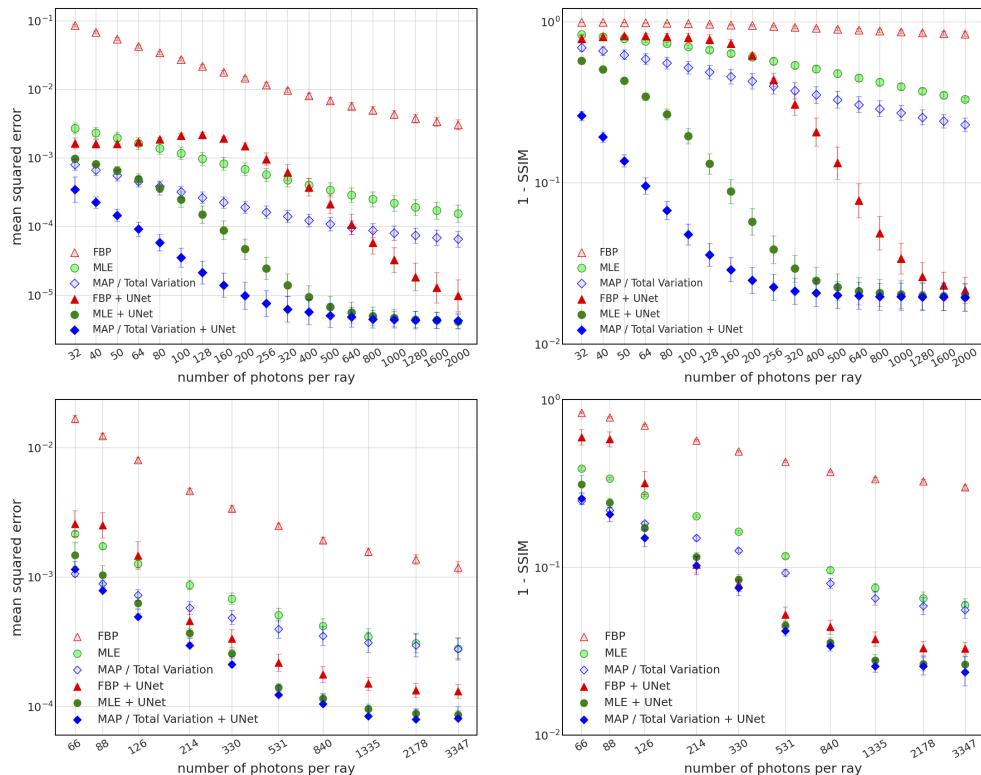| Method	Simulation (photons/ray)	Experiment (photons/ray)
FBP+UNet	640	531
MLE+UNet (+Gaussian)	128	330
MAP+UNet (+Gaussian+sparsity)	80	214

6. Discussion

In our study, we selected TV as the sparsity-promoting prior to the MAP and MAP+UNet algorithms. While TV is well-suited for reconstructing simple circuit structures, its applicability to more generalized imaging tasks with complex structures may be limited. Alternative sparsity-promoting priors, such as wavelet-based and Laplacian priors [77–79], could be explored to ensure the reconstruction performance of our method across various applications. In addition, we did not fine-tune the regularization parameter for TV regularized reconstructions under different noise conditions. Using the same value for all noise levels might be sub-optimal, suggesting room for further improvement.

Though the trends observed in the simulation are consistent with the experimental results, there is a discrepancy in the thresholds of acceptable performance. This can be attributed to several factors typically present in real-world experimental settings, such as noise sources, system imperfections, and simplified modeling.

7. Conclusions

We have demonstrated and evaluated quantitatively a noise-resilient deep-reconstruction approach for X-ray tomography with integrated circuits in both simulation and experiment. Using the maximum *a posteriori* reconstructions as the training inputs compensates the noise in the measurements, and leads to a learned prior that is noise-resilient. More importantly, such noise resilience is achieved without obtaining training samples at each possible noise distribution. The use of a sparsity-promoting prior is especially helpful for noisy data collected under low energy fluxes. Our noise-resilient deep-reconstruction algorithm may benefit applications with limited training sets due to long acquisition times and real-time dynamic imaging constrained by temporal change rates.

Appendix: quantitative results with alternative metrics

Fig. 8. Quantitative comparison between different reconstruction algorithms with mean squared error (MSE) and structural similarity index measure (SSIM) metrics. The top two figures are for simulated data, and the bottom two are for experimental results.

Here, we present our quantitative comparison of various reconstruction algorithms with mean squared error (MSE) and structural similarity index measure (SSIM) metrics. We find that our main conclusions do not depend on which metric is chosen.

References

1. R. A. Robb, *Three-dimensional Biomedical Imaging*, vol. 2 (CRC Press, 1985).
2. C. M. Tempany and B. J. McNeil, "Advances in biomedical imaging," JAMA **285**(5), 562–567 (2001).
3. R. Weissleder and M. Nahrendorf, "Advancing biomedical imaging," Proc. Natl. Acad. Sci. **112**(47), 14424–14428 (2015).
4. L. Salvo, P. Cloetens, E. Maire, S. Zabler, J. J. Blandin, J.-Y. Buffiere, W. Ludwig, E. Boller, D. Bellet, and C. Josserond, "X-ray micro-tomography an attractive characterisation technique in materials science," Nucl. Instrum. Methods Phys. Res., Sect. B **200**, 273–286 (2003).
5. G. Möbus and B. J. Inkson, "Nanoscale tomography in materials science," Mater. Today **10**(12), 18–25 (2007).
6. L. Salvo, M. Suéry, A. Marmottant, N. Limodin, and D. Bernard, "3D imaging in material science: Application of X-ray tomography," C. R. Phys. **11**(9-10), 641–649 (2010).
7. A. du Plessis, S. G. le Roux, G. Booysen, and J. Els, "Quality control of a laser additive manufactured medical implant by X-ray tomography," 3D Print. Addit. Manuf. **3**(3), 175–182 (2016).
8. A. Du Plessis, I. Yadroitsev, I. Yadroitsava, and S. G. Le Roux, "X-ray microcomputed tomography in additive manufacturing: a review of the current technology and applications," 3D Print. Addit. Manuf. **5**(3), 227–247 (2018).
9. H. Stark, J. Woods, I. Paul, and R. Hingorani, "Direct Fourier reconstruction in computer tomography," IEEE Trans. Acoust., Speech, Signal Process. **29**(2), 237–245 (1981).
10. A. K. Louis, "Incomplete data problems in X-ray computerized tomography," Numer. Math. **48**(3), 251–262 (1986).
11. M. E. Davison, "The ill-conditioned nature of the limited angle tomography problem," SIAM J. Appl. Math. **43**(2), 428–448 (1983).
12. G. Wang, Y. Zhang, X. Ye, and X. Mou, *Machine Learning for Tomographic Imaging* (IOP Publishing, 2019).
13. H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, vol. 375 (Springer Science & Business Media, 1996).
14. C. Bouman and K. Sauer, "A generalized Gaussian image model for edge-preserving MAP estimation," IEEE Trans. on Image Process. **2**(3), 296–310 (1993).
15. A. Chambolle and P.-L. Lions, "Image recovery via total variation minimization and related problems," Numer. Math. **76**(2), 167–188 (1997).
16. R. G. Baraniuk, "Compressive sensing [lecture notes]," IEEE Signal Process. Mag. **24**(4), 118–121 (2007).
17. L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," Phys. D **60**(1-4), 259–268 (1992).
18. V. Panin, G. Zeng, and G. Gullberg, "Total variation regulated EM algorithm," in *Nuclear Science Symposium and Medical Imaging Conference (Cat. No. 98CH36255)*, vol. 3 (IEEE, 1998), pp. 1562–1566.
19. D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via $\ell 1$ minimization," Proc. Natl. Acad. Sci. **100**(5), 2197–2202 (2003).
20. M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," IEEE Trans. on Image Process. **15**(12), 3736–3745 (2006).
21. K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proceedings of the 27th international conference on machine learning* (2010), pp. 399–406.
22. M. M. Lell and M. Kachelrieß, "Recent and upcoming technological developments in computed tomography: high speed, low dose, deep learning, multienergy," Invest. Radiol. **55**(1), 8–19 (2020).
23. J. Wang, J. Liang, J. Cheng, Y. Guo, and L. Zeng, "Deep learning based image reconstruction algorithm for limited-angle translational computed tomography," PLoS One **15**(1), e0226963 (2020).
24. T. Würfl, F. C. Ghesu, V. Christlein, and A. Maier, "Deep learning computed tomography," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, 2016), pp. 432–440.
25. T. Würfl, M. Hoffmann, V. Christlein, K. Breininger, Y. Huang, M. Unberath, and A. K. Maier, "Deep learning computed tomography: Learning projection-domain weights from image domain in limited angle problems," IEEE Trans. Med. Imaging **37**(6), 1454–1463 (2018).
26. Y. Huang, S. Wang, Y. Guan, and A. Maier, "Limited angle tomography for transmission X-ray microscopy using deep learning," J. Synchrotron Radiat. **27**(2), 477–485 (2020).
27. I. Kang, A. Goy, and G. Barbastathis, "Dynamical machine learning volumetric reconstruction of objects' interiors from limited angular views," Light: Sci. Appl. **10**(1), 1–21 (2021).
28. Z. Guo, J. K. Song, G. Barbastathis, M. E. Glinsky, C. T. Vaughan, K. W. Larson, B. K. Alpert, and Z. H. Levine, "Physics-assisted generative adversarial network for X-ray tomography," Opt. Express **30**(13), 23238–23259 (2022).

29. K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," IEEE Trans. on Image Process. **26**(9), 4509–4522 (2017).

30. Z. Li, S. Zhou, J. Huang, L. Yu, and M. Jin, "Investigation of low-dose CT image denoising using unpaired deep learning methods," IEEE Trans. Radiat. Plasma Med. Sci. **5**(2), 224–234 (2021).

31. I. Kang, Z. Wu, Y. Jiang, Y. Yao, J. Deng, J. Klug, S. Vogt, and G. Barbastathis, "Attentional ptycho-tomography (apt) for three-dimensional nanoscale x-ray imaging with minimal data acquisition and computation time," arXiv, arXiv:2212.00014 (2022).

32. M. Sugiyama and M. Kawanabe, *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation* (MIT Press, 2012).

33. K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang, "Domain adaptation under target and conditional shift," in *International Conference on Machine Learning* (PMLR, 2013), pp. 819–827.

34. V. Antun, F. Renna, C. Poon, B. Adcock, and A. C. Hansen, "On instabilities of deep learning in image reconstruction and the potential costs of AI," Proc. Natl. Acad. Sci. **117**(48), 30088–30095 (2020).

35. I. Tunali, L. O. Hall, S. Napel, D. Cherezov, A. Guvenis, R. J. Gillies, and M. B. Schabath, "Stability and reproducibility of computed tomography radiomic features extracted from peritumoral regions of lung cancer lesions," Med. Phys. **46**(11), 5075–5085 (2019).

36. W. Wu, D. Hu, W. Cong, H. Shan, S. Wang, C. Niu, P. Yan, H. Yu, V. Vardhanabhuti, and G. Wang, "Stabilizing deep tomographic reconstruction networks," arXiv, arXiv:2008.01846 (2020).

37. C. D. Pain, G. F. Egan, and Z. Chen, "Deep learning-based image reconstruction and post-processing methods in positron emission tomography for low-dose imaging and resolution enhancement," Eur. J. Nucl. Med. Mol. Imaging **49**(9), 3098–3118 (2022).

38. "Rapid analysis of various emerging nanoelectronics," https://www.iarpa.gov/research-programs/raven. Accessed: 2023-03-27.

39. Z. H. Levine, A. R. Kalukin, S. P. Frigo, I. McNulty, and M. Kuhn, "Tomographic reconstruction of an integrated circuit interconnect," Appl. Phys. Lett. **74**(1), 150–152 (1999).

40. M. Holler, M. Guizar-Sicairos, E. H. Tsai, R. Dinapoli, E. Müller, O. Bunk, J. Raabe, and G. Aeppli, "High-resolution non-destructive three-dimensional imaging of integrated circuits," Nature **543**(7645), 402–406 (2017).

41. Z. H. Levine, B. K. Alpert, A. L. Dagel, J. W. Fowler, E. S. Jimenez, N. Nakamura, D. S. Swetz, P. Szypryt, K. R. Thompson, and J. N. Ullom, "A tabletop x-ray tomography instrument for nanometer-scale imaging: Reconstructions," Microsystems & Nanoengineering, in press (2023).

42. Z. Guo, "Noise resilience deep reconstruction for X-ray tomography," Github (2022), https://github.com/zguo0525/Noise-resilience-deep-reconstruction-for-X-ray-Tomography.

43. Y. Wei, G. Wang, and J. Hsieh, "An intuitive discussion on the ideal ramp filter in computed tomography (i)," Comput. & Math. with Appl. **49**(5-6), 731–740 (2005).

44. D. E. Dudgeon and R. M. Mersereau, *Multidimensional Digital Signal Processing* (Prentice-Hall, 1984).

45. F. E. Boas and D. Fleischmann, "Evaluation of two iterative techniques for reducing metal artifacts in computed tomography," Radiology **259**(3), 894–902 (2011).

46. T.-B. Chen, H.-Y. Chen, M.-C. Lin, L.-W. Lin, N.-H. Lu, F.-S. Tsai, and Y.-H. Huang, "Delimited strike artifacts from FBP using a robust morphological structure operation," Radiat. Phys. Chem. **97**, 31–37 (2014).

47. T. Kailath, "Lectures on Wiener and Kalman filtering," in *Lectures on Wiener and Kalman Filtering*, (Springer, 1981), pp. 1–143.

48. G. H. Golub, P. C. Hansen, and D. P. O'Leary, "Tikhonov regularization and total least squares," SIAM J. Matrix Anal. & Appl. **21**(1), 185–194 (1999).

49. G. Wang, T. Frei, and M. W. Vannier, "Fast iterative algorithm for metal artifact reduction in X-ray CT," Acad. radiology **7**(8), 607–614 (2000).

50. M. Beister, D. Kolditz, and W. A. Kalender, "Iterative reconstruction methods in X-ray CT," Phys. Medica **28**(2), 94–108 (2012).

51. I. A. Elbakri and J. A. Fessler, "Statistical image reconstruction for polyenergetic X-ray computed tomography," IEEE Trans. Med. Imaging **21**(2), 89–99 (2002).

52. Z. H. Levine, A. J. Kearsley, and J. G. Hagedorn, "Bayesian tomography for projections with an arbitrary transmission function with an application in electron microscopy," J. Res. Natl. Inst. Stand. Technol. **111**(6), 411–417 (2006).

53. N. P. Galatsanos and A. K. Katsaggelos, "Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation," IEEE Trans. on Image Process. **1**(3), 322–336 (1992).

54. M. Benning and M. Burger, "Modern regularization methods for inverse problems," Acta Numer. **27**, 1–111 (2018).

55. E. Kobler, M. Muckley, B. Chen, F. Knoll, K. Hammernik, T. Pock, D. Sodickson, and R. Otazo, "Variational deep learning for low-dose computed tomography," in *International Conference on Acoustics, Speech and Signal Processing* (IEEE, 2018), pp. 6687–6691.

56. M. Araya-Polo, J. Jennings, A. Adler, and T. Dahlke, "Deep-learning tomography," The Leading Edge **37**(1), 58–66 (2018).

57. T. A. Bubba, G. Kutyniok, M. Lassas, M. März, W. Samek, S. Siltanen, and V. Srinivasan, "Learning the invisible: A hybrid deep learning-shearlet framework for limited angle computed tomography," Inverse Probl. **35**(6), 064002 (2019).

58. Y. Huang, A. Preuhs, G. Lauritsch, M. Manhart, X. Huang, and A. Maier, "Data consistent artifact reduction for limited angle tomography with deep learning prior," in *International Workshop on Machine Learning for Medical Image Reconstruction*, (Springer, 2019), pp. 101–112.

59. W. Wang, X.-G. Xia, C. He, Z. Ren, J. Lu, T. Wang, and B. Lei, "An end-to-end deep network for reconstructing ct images directly from sparse sinograms," IEEE Trans. Comput. Imaging **6**, 1548–1560 (2020).

60. J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "Generative adversarial networks for noise reduction in low-dose ct," IEEE Trans. Med. Imaging **36**(12), 2536–2545 (2017).

61. H. Chen, Y. Zhang, Y. Chen, J. Zhang, W. Zhang, H. Sun, Y. Lv, P. Liao, J. Zhou, and G. Wang, "LEARN: learned experts' assessment-based reconstruction network for sparse-data CT," IEEE Trans. Med. Imaging **37**(6), 1333–1347 (2018).

62. I. Y. Chun, Z. Huang, H. Lim, and J. Fessler, "Momentum-net: Fast and convergent iterative neural network for inverse problems," IEEE Trans. Pattern Anal. Mach. Intell. **45**(4), 4915–4931 (2023).

63. D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, D. P. Naidich, and S. Shetty, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," Nat. Med. **25**(6), 954–961 (2019).

64. J. Dong, J. Fu, and Z. He, "A deep learning reconstruction framework for X-ray computed tomography with incomplete data," PLoS One **14**(11), e0224426 (2019).

65. K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," J. Big Data **3**(1), 9–40 (2016).

66. A. Reuther, J. Kepner, C. Byun, S. Samsi, W. Arcand, D. Bestor, B. Bergeron, V. Gadepally, M. Houle, M. Hubbell, M. Jones, A. Klein, L. Milechin, J. Mullen, A. Prout, A. Rosa, C. Yee, and P. Michaleas, "Interactive supercomputing on 40,000 cores for machine learning and data analysis," in *High Performance Extreme Computing Conference* (IEEE, 2018), pp. 1–6.

67. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," arXiv, arXiv:1711.05101 (2017).

68. I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," arXiv, arXiv:1608.03983 (2016).

69. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv, arXiv:1412.6980 (2014).

70. W. Van Aarle, W. J. Palenstijn, J. Cant, E. Janssens, F. Bleichrodt, A. Dabravolski, J. De Beenhouwer, K. J. Batenburg, and J. Sijbers, "Fast and flexible X-ray tomography using the ASTRA toolbox," Opt. Express **24**(22), 25129–25147 (2016).

71. P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: appropriate use and interpretation," Anesth. & Analg. **126**(5), 1763–1768 (2018).

72. K. Yen, E. K. Yen, and R. G. Johnston, "The ineffectiveness of the correlation coefficient for image comparisons," Los Alamos National Laboratory report LA-UR-96-2474 (1996).

73. S. Mallat, "Group invariant scattering," Comm. Pure Appl. Math. **65**(10), 1331–1398 (2012).

74. J. Bruna and S. Mallat, "Invariant scattering convolution networks," IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1872–1886 (2013).

75. M. Andreux, T. Angles, G. Exarchakis, R. Leonarduzzi, G. Rochette, L. Thiry, J. Zarka, S. Mallat, J. Andén, E. Belilovsky, J. Bruna, V. Lostanlen, M. Chaudhary, M. J. Hirn, E. Oyallon, S. Zhang, C. Cella, and M. Eickenberg, "Kymatio: Scattering transforms in python," J. Machine Learning Research **21**(60), 1–6 (2020).

76. A. Bernardino and J. Santos-Victor, "A real-time gabor primal sketch for visual attention," in *Iberian Conference on Pattern Recognition and Image Analysis* (Springer, 2005), pp. 335–342.

77. I. Loris, G. Nolet, I. Daubechies, and F. Dahlen, "Tomographic inversion using l1-norm regularization of wavelet coefficients," Geophys. J. Int. **170**(1), 359–370 (2007).

78. A. Borsdorf, R. Raupach, T. Flohr, and J. Hornegger, "Wavelet based noise reduction in CT-images using correlation analysis," IEEE Trans. Med. Imaging **27**(12), 1685–1703 (2008).

79. H. Antil, Z. W. Di, and R. Khatri, "Bilevel optimization, deep learning and fractional laplacian regularization with applications in tomography," Inverse Probl. **36**(6), 064001 (2020).