Equivariance2Inverse: A Practical Self-Supervised CT Reconstruction Method Benchmarked on Real, Limited-Angle, and Blurred Data

Dirk Elias Schut[1], Adriaan Graas[1], Robert van Liere[1,2], Tristan van Leeuwen[1,3]

***Abstract*—Deep learning has shown impressive results in reducing noise and artifacts in X-ray computed tomography (CT) reconstruction. Self-supervised CT reconstruction methods are especially appealing for real-world applications because they require no ground truth training examples. However, these methods involve a simplified X-ray physics model during training, which may make inaccurate assumptions, for example, about scintillator blurring, the scanning geometry, or the distribution of the noise. As a result, they can be less robust to real-world imaging circumstances. In this paper, we review the model assumptions of six recent self-supervised CT reconstruction methods. Based on this, we combined concepts of the Robust Equivariant Imaging and Sparse2Inverse methods in a new self-supervised CT reconstruction method called Equivariance2Inverse that is robust to scintillator blurring and limited-angle data. We benchmarked Equivariance2Inverse and the existing methods on the real-world 2DeteCT dataset and on synthetic data with and without scintillator blurring and a limited-angle scanning geometry. The results of our benchmark show that methods that assume that the noise is pixel-wise independent do not perform well on data with scintillator blurring. Moreover, they show that when the distribution of objects is rotationally invariant, this invariance can be used to reduce artifacts in limited-angle reconstructions.**

I. Introduction

In X-ray computed tomography (CT), multiple X-ray projection images are combined to form an image representing the inside of an object through a process called image reconstruction. Learned image reconstruction techniques have shown impressive results in reducing noise and artifacts compared to traditional (non-learned) image reconstruction techniques [1], [2]. This is very promising for low-dose (e.g., medical) or high-throughput (e.g., industrial) applications of CT imaging. Learned image reconstruction was first demonstrated using supervised learning. However, supervised learning requires a large dataset of paired input and ground truth data, which can be challenging or expensive to acquire. Unsupervised CT reconstruction methods do not require paired input and ground truth data [3], making these methods more practical for real-world use.

Several approaches for unsupervised CT reconstruction exist that use different data and training strategies. Diffusion-based methods [4]–[7] learn a prior distribution of the reconstructed volumes, and they have outperformed supervised learning-based methods for several reconstruction problems. However, diffusion-based methods require ground truth data of objects for training, which still makes it challenging to acquire a suitable training dataset. Methods based on implicit neural representations (INRs) [8]–[11] train a separate neural network for each scan. INR-based methods are particularly useful when only a single scan of an object is available; however, they are less suitable for high-throughput applications, as training a new network for every scan is computationally intensive. Moreover, these methods do not benefit from the large diversity of image features that large datasets have to offer. INR-based methods are sometimes referred to as self-supervised. However, we will use the term self-supervised exclusively to refer to a different category of methods. Self-supervised methods are trained on a dataset of measurement data from multiple scans, where in every loss function call, data from the same scan is used both as input and as target [12]–[18]. Measurement data is typically simpler to acquire than ground truth data, making self-supervised methods simpler to train in practice than diffusion-based methods, while offering better performance than INR-based methods. Therefore, this paper focuses on self-supervised methods.

To train a neural network without ground truth data, self-supervised methods rely on an X-ray physics model. Traditional CT reconstruction methods also use an X-ray physics model, which is often highly simplified, for example, assuming a dense-view geometry, a linear projection operator, and additive Gaussian noise [19]. Recent self-supervised CT reconstruction methods have introduced different assumptions, such as a sparse-view geometry [13], [15], [17], [18], a non-linear projection operator with Poisson + Gaussian noise [13], and correlated noise [16], [20]. The fact that different reconstruction methods make different assumptions raises the question of how well these assumptions reflect real-world data. When the same model assumptions are used for generating data and for evaluating a reconstruction method on that synthetic data, the results may be unrealistically positive. This is known as an inverse crime [21], [22]

In this paper, we investigate how limited-angle geometries [23] and scintillator blurring [24] affect self-supervised CT reconstruction. These conditions are common in practice, but they are commonly not considered when developing self-supervised CT reconstruction methods. We review the model assumptions of six recent self-supervised CT reconstruction methods. Based on this, we combine concepts of the Robust Equivariant Imaging [13] and Sparse2Inverse [15] methods in a new self-supervised CT reconstruction method called Equivariance2Inverse that is robust to scintillator blurring and limited-angle data. Moreover, we benchmark Equivariance2Inverse and the six recent methods to evaluate how their

[1] Computational Imaging Group, Centrum Wiskunde en Informatica (CWI), Science Park 123, 1098 XG Amsterdam, The Netherlands
[2] Visualization Cluster, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands
[3] Mathematisch Instituut, Utrecht University, Budapestlaan 6, 3584 CD Utrecht, The Netherlands

This article has been accepted for publication in IEEE Transactions on Computational Imaging. This version includes all revisions by the authors based on peer reviewer suggestions, but it has not been fully edited, and content may change prior to final publication. Citation information: DOI 10.1109/TCI.2026.3684416

model assumptions affect the reconstruction performance. For this goal, synthetic and real-world data have complementary strengths. Synthetic data can be generated with any X-ray physics model, making it possible to change the model assumptions in isolation [25], [26]. Real-world data provides a good indication of how a method will perform in practice. Our benchmark uses synthetic data with and without scintillator blurring and a limited-angle geometry to test the robustness of each method to these effects. Moreover, our benchmark uses two datasets of the real-world 2DeteCT dataset [27].

The structure of the paper is as follows: Section II outlines the six recent self-supervised CT reconstruction methods. It first describes common assumptions on X-ray physics, and then describes the different approaches used for self-supervised training. Section III presents our novel self-supervised CT reconstruction method Equivariance2Inverse (E2I). Section IV describes our benchmark and section V describes an ablation study. Section VI describes the results of the benchmark and the ablation study and relates them to the model assumptions of each method. Finally, Sections VII and VIII are the discussion and conclusion.

II. Background

A. Problem formulation

A CT scanner collects multiple X-ray projection images of an object, and the CT reconstruction algorithm uses this data to create an image of the X-ray attenuation coefficient inside that object. In this paper, 2D objects and 1D detectors are considered. When n X-ray projection images are acquired with a detector that has m pixels, all measurements can be represented by a vector $\boldsymbol{y} \in \mathbb{R}^{nm}$. The attenuation coefficient inside the object is discretized into a grid of j by k pixels, represented by a vector $\boldsymbol{x} \in \mathbb{R}^{jk}$. A CT reconstruction algorithm is a function $f : \mathbb{R}^{nm} \to \mathbb{R}^{jk}$, and it performs the task of deriving $\boldsymbol{x}$ from the X-ray images $\boldsymbol{y}$. However, there may be multiple objects $\boldsymbol{x}$ that produce the same measurements $\boldsymbol{y}$ because of noise or incomplete measurements. This can be modeled with random variables (which will be notated as bold capital letters): $\boldsymbol{X}$ for the objects, and $\boldsymbol{Y}$ for the projection data. For a given loss function $l(\cdot)$ the reconstruction method aims to minimize:

$$\hat{f} = \underset{f}{\operatorname{argmin}} \left(\mathbb{E} \left[l(f(\boldsymbol{Y}), \boldsymbol{X}) \right] \right). \tag{1}$$

All expected values in this paper are calculated over all random variables. The joint distribution between $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be decomposed into the conditional distribution $p(\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x})$, and the prior distribution of $\boldsymbol{X}$.

B. Forward models

A forward model approximates the conditional distribution $p(\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x})$ by modeling the X-ray physics. All self-supervised methods rely on assumptions related to the forward model, but the assumptions vary between methods, and will be discussed in Section II-D. A forward model can also be used to generate synthetic CT projection data.

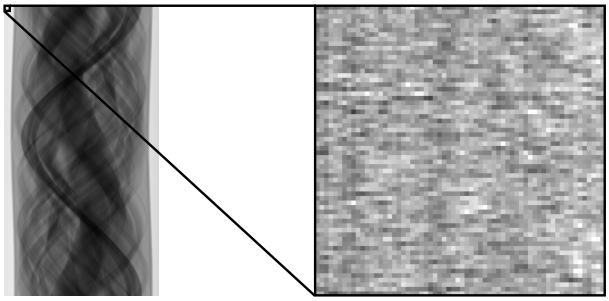

Fig. 1. Example of scintillator blur in the 2DeteCT dataset. Each horizontal line corresponds to a detector readout at a given time. The zoomed-in region corresponds to constant background radiation, so all variations over time are due to noise. The horizontal correlations in the noise can be attributed to scintillator blurring.

1) X-Ray physics: Here we will explain several aspects of X-ray physics and combine them into a forward model. This model will be used for generating the synthetic benchmark data:

$$\begin{aligned} \boldsymbol{P} &\sim \text{Poisson}\left(\text{diag}(\boldsymbol{c}) \exp\left(-A\boldsymbol{x}\right)\right) \\ \boldsymbol{G} &\sim \text{Gaussian}\left(\boldsymbol{u}, \text{diag}(\boldsymbol{v})\right) \\ \boldsymbol{Y} &= \text{diag}(\boldsymbol{w}) B \boldsymbol{P} + \boldsymbol{G}. \end{aligned} \tag{2}$$

X-rays are emitted by an X-ray source and they decay exponentially, depending on the local attenuation coefficient of the object $\boldsymbol{x}$ they are propagating through. The linear projection operator $A \in \mathbb{R}^{nm \times jk}$ describes how the X-rays traverse through the object, so $\exp\left(-A\boldsymbol{x}\right)$ is the absorption for each detector pixel in each projection image. The number of X-ray photons reaching the detector is modeled as a Poisson-distributed random vector $\boldsymbol{P}$, due to the quantum nature of X-rays [19]. The mean photon count without attenuation $\boldsymbol{c} \in \mathbb{R}^{jk}$ is direction-dependent in cone beam CT scanners, because of the anode heel effect [28].

Energy-integrating X-ray detectors consist of a scintillator and a sensor layer. The scintillator converts each X-ray photon into multiple visible light photons, and the sensor layer measures the visible light. The conversion from X-ray photons to detector counts can be characterized by a gain $\boldsymbol{w} \in \mathbb{R}^{nm}$, which is pixel-dependent [26]. The visible light photons scatter in the scintillator before reaching the sensor layer, resulting in blurring [19], [29], [30]. This scintillator blurring can be approximated as a convolution $B \in \mathbb{R}^{nm \times nm}$. Scintillator blurring not only blurs the signal, but also the Poisson component of the noise, resulting in correlated noise (e.g. Figure 1). Moreover, the electronics in visible light sensors introduce Gaussian noise $\boldsymbol{G}$ into their measurements with a pixel-wise variance ($\boldsymbol{v} \in \mathbb{R}^{nm}$), and even without any X-rays, there may be a small signal $\boldsymbol{u} \in \mathbb{R}^{jk}$ [31].

2) Pre-processing: In many CT reconstruction methods, pre-processing (also called flatfielding [32] and log-transforming) is applied to the raw data $\boldsymbol{Y}$ to obtain the pre-processed data $\widetilde{\boldsymbol{Y}}$:

$$\widetilde{\boldsymbol{Y}} = -\log\left(\text{diag}(\boldsymbol{p} - \boldsymbol{q})^{-1}(\boldsymbol{Y} - \boldsymbol{q})\right). \tag{3}$$

The values of $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^{nm}$ are obtained using simple calibration measurements. $\boldsymbol{p}$ is obtained by averaging multiple

measurements with no object in the scanner, and it roughly corresponds to $\boldsymbol{c} \odot \boldsymbol{w} + \boldsymbol{u}$ (where $\odot$ is the element-wise product) in Equation 2. $\boldsymbol{q}$ is obtained by averaging multiple measurements with the X-ray source turned off, and it roughly corresponds to $\boldsymbol{u}$ in Equation 2. For $\widetilde{\boldsymbol{Y}}$ a linear forward model with additive Gaussian noise with covariance $\Sigma \in \mathbb{R}^{nm \times nm}$ is commonly assumed [19], [33]:

$$\widetilde{\boldsymbol{Y}} \sim \text{Gaussian}\left(A\boldsymbol{x}, \Sigma\right). \tag{4}$$

3) The projection operator: The projection operator A is determined by the scanning geometry, and by the number of acquired projection images. The scanning geometry is how the source, detector, and object move relative to each other, and it is called *complete* when the ranges of motion are sufficient for reconstructing any object within the volume of interest [34], [35]. Circular geometries that cover an insufficient range of rotations to be complete are called *limited-angle* geometries. Reconstructions from incomplete geometries may have blurring or streaking artifacts. Blurring and streaking artifacts may also appear when the number of projection images is low. Such reconstruction problems are called sparse-view reconstruction problems [15], [17], [36].

C. Supervised CT reconstruction (SUP)

Equation 1 can be interpreted as a supervised deep learning problem. In that setting, the optimization of the reconstruction function f is performed by drawing paired samples from $(\boldsymbol{X}, \boldsymbol{Y})$, and doing stochastic gradient descent over these samples, which approximates optimizing over the expected value of the loss.

In this paper, we follow the FBPConvNet approach [1], where a neural network $g : \mathbb{R}^{jk} \to \mathbb{R}^{jk}$ is applied as a post-process to a Filtered Backprojection (FBP) reconstruction [33]. An FBP reconstruction can be represented by a matrix $R \in \mathbb{R}^{jk \times nm}$, and it requires pre-processed projection data $\widetilde{\boldsymbol{Y}}$ as input. Together g and R form a learned reconstruction function for pre-processed data: $g(R(\widetilde{\boldsymbol{Y}}))$. A mean squared error (MSE) loss is used to optimize the parameters of g, resulting in the loss function:

$$\mathbb{E}\left[\left\|g(R\widetilde{\boldsymbol{Y}}) - \boldsymbol{X}\right\|_2^2\right]. \tag{5}$$

D. Self-supervised CT reconstruction methods

In this section, we review six recent self-supervised CT reconstruction methods. Unless otherwise mentioned, pre-processed projection data was used as input ($\widetilde{\boldsymbol{Y}}$ in Equations 3 & 4). The loss functions and the assumptions made by each method are provided in Table I, and an illustration of how the methods are calculated is provided in Figure 2.

1) Cross-validation methods: Cross-validation methods split the projection data into two parts: *the network input data* and *the target data*. The network is trained to predict the target data from the network input data. Different splits are used in different training iterations so that all data is assigned both as network input and as target data. Cross-validation methods require that the noise is independent and zero mean between both parts. The network can learn to approximate the signal of the target data because this information is correlated to the network input data, but it can not learn to predict the noise of the target data because the noise is independent. Therefore, it should converge towards the minimum mean squared error (MMSE) estimator of the noise-free target data from the noisy input data [37], [38]. After the network has finished training, one approach for inference is to average the reconstructions from multiple input splits [12]. Another approach is to use all data as input during inference. While this approach may introduce some bias, it has shown good experimental performance [14], [15], [39], and is less computationally expensive. This second approach was used for all cross-validation methods in the benchmark of this paper.

The main benefits of cross-validation methods are that their required assumptions are often met in practice, and that they do not require knowing the exact noise level. However, a downside is that a more efficient or less biased estimator may be possible because the full data can not be used as the network input data during training.

We compare three cross-validation CT reconstruction methods with slightly different loss functions (see Table I): In **Noise2Inverse (N2I)** [12] 25% equally spaced projection images are used as target data, and the remaining projection images are used as input data. The neural network weights are optimized to minimize the MSE between the neural network output and an FBP reconstruction of the target data. **Sparse2Inverse (S2I)** [15] uses the same splits between target and network input data as N2I. However, instead of performing an FBP reconstruction of the target data, the neural network output is projected using matrix A, and the MSE is calculated between the projected neural network output and the target data. **Proj2Proj (P2P)** [14] uses pixel-wise instead of projection-wise splitting between network input and target data. The network input data is the projection data with every fourth pixel horizontally and vertically replaced by its local mean. The loss is calculated in the projection domain, like S2I, but only over the pixels that were replaced in the neural network input.

For sparse-view or limited-angle reconstruction problems, N2I may learn to approximate streaking artifacts, because the FBP reconstructions of the target data contain streaking artifacts. S2I was designed to avoid this problem by not performing an FBP reconstruction of the target data. While this approach does not incentivize learning streaking artifacts, the neural network may learn to produce arbitrary components in the null-space of A, because adding any null-space component to the neural network output does not affect the loss. Nevertheless, in the experiments of the original S2I paper, S2I consistently outperformed N2I on sparse-view data [15].

Scintillator blurring was not mentioned in the original publications of any of these methods. However, it is expected that blurring will negatively affect the denoising performance of P2P, because blurring introduces correlations in the noise between neighboring pixels, which for P2P violates the requirement that the noise should be independent between the network input and target data. Blurring does not cause correlations between projections, so N2I and S2I should be relatively

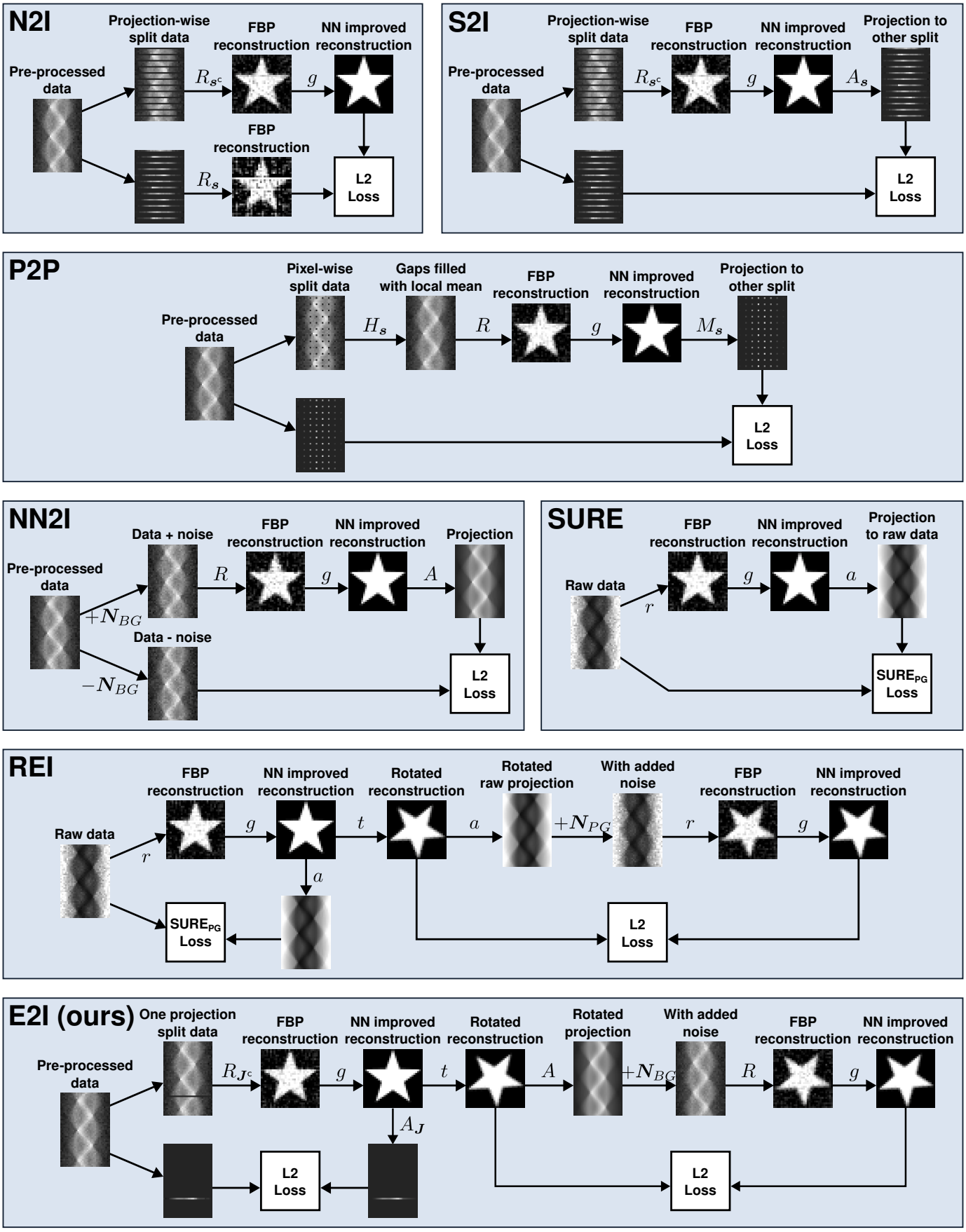

Fig. 2. Illustration of how the loss is calculated in the self-supervised CT reconstruction methods. The arrows with letters correspond to function calls or matrix multiplications using the same notation as in Table I.

TABLE I
LOSS FUNCTIONS AND ASSUMPTIONS OF THE COMPARED SELF-SUPERVISED CT RECONSTRUCTION METHODS.

Method	Loss Function	Assumptions
Noise2Inverse (N2I)	$\mathbb{E}\left[\frac{1}{4}\sum_{\boldsymbol{s}\in\{\boldsymbol{s}_1,..,\boldsymbol{s}_4\}}\left\|g(R_{\boldsymbol{s}^c}\widetilde{\boldsymbol{Y}}_{\boldsymbol{s}^c})-R_{\boldsymbol{s}}\widetilde{\boldsymbol{Y}}_{\boldsymbol{s}}\right\|_2^2\right]$	Pre-processed noise is projection-wise independent and zero mean
Sparse2Inverse (S2I)	$\mathbb{E}\left[\frac{1}{4}\sum_{\boldsymbol{s}\in\{\boldsymbol{s}_1,..,\boldsymbol{s}_4\}}\left\|A_{\boldsymbol{s}}g(R_{\boldsymbol{s}^c}\widetilde{\boldsymbol{Y}}_{\boldsymbol{s}^c})-\widetilde{\boldsymbol{Y}}_{\boldsymbol{s}}\right\|_2^2\right]$	Pre-processed noise is projection-wise independent and zero mean
Proj2Proj (P2P)	$\mathbb{E}\left[\frac{1}{16}\sum_{\boldsymbol{s}\in\{\boldsymbol{s}_1,..,\boldsymbol{s}_{16}\}}\left\|M_{\boldsymbol{s}}Ag(RH_{\boldsymbol{s}}\widetilde{\boldsymbol{Y}})-M_{\boldsymbol{s}}\widetilde{\boldsymbol{Y}}\right\|_2^2\right]$	Pre-processed noise is pixel-wise independent and zero mean
Noisier2Inverse (NN2I)	$\mathbb{E}\left[\left\|Ag(R(\widetilde{\boldsymbol{Y}}+\boldsymbol{N}_{\text{BG}}))-(\widetilde{\boldsymbol{Y}}-\boldsymbol{N}_{\text{BG}})\right\|_2^2\right]$	Pre-processed noise is blurred Gaussian with known parameters
SURE	$\mathbb{E}[\text{SURE}_{\text{PG}}(a(g(r(\boldsymbol{Y}))),\boldsymbol{Y})]$	Raw noise is Poisson + Gaussian with known parameters
Robust Equivariant Imaging (REI)	$\mathbb{E}[\text{SURE}_{\text{PG}}(a(g(r(\boldsymbol{Y}))),\boldsymbol{Y})+\lambda h_{\text{REI}}(\boldsymbol{Y})]$	Raw noise is Poisson + Gaussian with known parameters and $\boldsymbol{X}$ is rotation invariant
Equivariance2Inverse (E2I)	$\mathbb{E}\left[\left\|A_{\boldsymbol{J}}(g(R_{\boldsymbol{J}^c}(\widetilde{\boldsymbol{Y}}_{\boldsymbol{J}^c})))-\widetilde{\boldsymbol{Y}}_{\boldsymbol{J}}\right\|_2^2+\lambda h_{\text{E2I}}(\widetilde{\boldsymbol{Y}}_{\boldsymbol{J}^c},\boldsymbol{J}^c)\right]$	Pre-processed noise is blurred Gaussian with known parameters and $\boldsymbol{X}$ is rotation invariant

g is the neural network that is optimized. A is the projection operator, and R is an FBP reconstruction. $a(\cdot)$ and $r(\cdot)$ are non-linear versions of A and R including the (inverse) pre-processing of Equation 3. $\boldsymbol{Y}$ is the raw projection data, and $\widetilde{\boldsymbol{Y}}$ is the pre-processed projection data. In N2I, S2I and E2I, $\widetilde{\boldsymbol{Y}}_{\boldsymbol{s}}$ are only the projection images with indices in $\boldsymbol{s}$, and $\widetilde{\boldsymbol{Y}}_{\boldsymbol{s}^c}$ are all other projection images in $\widetilde{\boldsymbol{Y}}$. In P2P the sets $\boldsymbol{s}$ represent pixelwise selections, and $H_{\boldsymbol{s}}$ is the operator that replaces the pixels in set $\boldsymbol{s}$ with their local means and $M_{\boldsymbol{s}}$ is a mask that only selects the pixels in $\boldsymbol{s}$. $\boldsymbol{N}_{\text{BG}}$ is blurred Gaussian noise that is randomly generated every time the loss function is calculated. SURE_{PG} is the SURE loss for a Poisson + Gaussian noise distribution. h_{REI} and h_{E2I} are equivariance terms and their definitions can be found in Equations 7 and 9, respectively. $\boldsymbol{J}$ is the index of one projection image, randomly sampled every time the loss function is calculated.

unaffected. This may explain why N2I was not affected by scintillator blurring when applied to real data [39], while pixel-wise splitting for X-ray denoising was affected [24].

2) Noisier2Inverse (NN2I): In NN2I [16], new noise is generated from a blurred Gaussian distribution, which should approximate the distribution of the noise in the projection data, and the neural network is trained to remove the noise. During training, the neural network is applied to the projection data with new noise added, and an MSE loss is calculated between this value and the projection data with the same noise subtracted (see Table I). This converges to the MMSE estimator of the noise-free projection data from the noisy projection data with additional added noise ($\widetilde{\boldsymbol{Y}}+\boldsymbol{N}_{\text{BG}}$). During inference, the network is applied to projection data with no additional added noise, leading to some bias, but producing good results in practice [16].

The main benefit of NN2I is that it is the only method in this section that was designed and tested for cases where correlated noise is present. When correlated noise is assumed, the added noise should simply be correlated in the same way. A downside of this is that it requires estimating the noise correlation and the noise level.

3) Stein's Unbiased Risk Estimator (SURE): SURE [40] is a function that uses knowledge of the noise model to provide an unbiased estimator of the MSE. Variants of SURE exist for multiple noise models [20], [41]. However, to the knowledge of the authors, no SURE-based estimator has currently been derived that matches the physics-based forward model of Equation 2.

A variant of SURE exists for Poisson + Gaussian noise with uniform gain $\gamma \in \mathbb{R}$ and standard deviation $\sigma \in \mathbb{R}$ [42]. For a given vector of noise-free projection data $\boldsymbol{z} \in \mathbb{R}^{jk}$, the noisy measurement is assumed to be a random variable $\boldsymbol{Z} = \gamma\boldsymbol{P}+\boldsymbol{G}$, with $\boldsymbol{P} \sim \text{Poisson}(\boldsymbol{z}/\gamma)$ and $\boldsymbol{G} \sim \text{Gaussian}\left(\boldsymbol{0}, \sigma^2 I\right)$. Moreover, let $b : \mathbb{R}^{jk} \to \mathbb{R}^{jk}$ be a weakly differentiable function. In that case the expectation of the SURE loss equals the expectation of the MSE between the processed noisy measurements $b(\boldsymbol{Z})$ and the noisy-free data $\boldsymbol{z}$:

$$\mathbb{E}\left[\text{SURE}_{\text{PG}}(b(\boldsymbol{Z}),\boldsymbol{Z})\right] = \mathbb{E}\left[\left\|b(\boldsymbol{Z})-\boldsymbol{z}\right\|_2^2\right]. \tag{6}$$

This SURE_{PG} loss was used for self-supervised CT reconstruction [13], and we refer to that paper on how to calculate SURE_{PG}. The Poisson + Gaussian assumption pertains to the raw data, whereas the intermediate FBP reconstruction requires pre-processed data. To make this work together, the FBP reconstruction operator R was replaced with a non-linear reconstruction operator $r : \mathbb{R}^{nm} \to \mathbb{R}^{jk}$, which includes the pre-processing done in Equation 3, and the forward operator A was replaced with a non-linear forward operator $a : \mathbb{R}^{jk} \to \mathbb{R}^{nm}$, which includes the inverse of the pre-processing.

The main benefit of SURE is that it converges towards the MMSE estimator of the noise-free projection data from the noisy projection data, which is very similar to the supervised learning loss. The main downside is that SURE requires modeling of the full forward model and calibration of its model parameters. SURE can be sensitive to calibration errors [20].

4) Robust Equivariant Imaging (REI): The distribution of $\boldsymbol{X}$ often contains the same object in multiple orientations. The distribution of $\boldsymbol{X}$ is said to be invariant to rotations if for every $\boldsymbol{x}$ in the distribution of $\boldsymbol{X}$ and every rotation matrix $Q \in \mathbb{R}^{jk\times jk}$, the probabilities of $\boldsymbol{x}$ and $Q\boldsymbol{x}$ are equal: $\mathbb{P}(\boldsymbol{x}) = \mathbb{P}(Q\boldsymbol{x})$. REI [13] optimizes a loss consisting of the SURE_{PG} loss (Equation 6), and an additional equivariance

term $\mathbb{E}\left[\lambda h_{\text{REI}}\left(\boldsymbol{Y}\right)\right]$:

$$\begin{aligned} \widetilde{\boldsymbol{X}}_1 &= t(g(r(\boldsymbol{Y})), \boldsymbol{T}) \\ \widetilde{\boldsymbol{X}}_2 &= g(r(a(\widetilde{\boldsymbol{X}}_1) + \boldsymbol{N}_{\text{PG}})) \\ \lambda h_{\text{REI}}\left(\boldsymbol{Y}\right) &= \lambda \left\| \widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_2 \right\|_2^2 \end{aligned} \tag{7}$$

$g(r(\boldsymbol{Y}))$ is the reconstructed image (where r is the non-linear version of R). Function $t(\cdot)$ rotates this image by a random amount $\boldsymbol{T}$. New projection data is generated from the rotated image by applying the projection operator a and adding noise $\boldsymbol{N}_{\text{PG}}$ with a Poisson + Gaussian distribution, which is assumed to be the distribution of the noise in $\boldsymbol{Y}$. From this projection data a new image $\widetilde{\boldsymbol{X}}_2$ is reconstructed. Because $\widetilde{\boldsymbol{X}}_1$ was rotated, its sparse-view and limited-angle artifacts should be in different positions than in $\widetilde{\boldsymbol{X}}_2$. Therefore, optimizing over the MSE between $\widetilde{\boldsymbol{X}}_1$ and $\widetilde{\boldsymbol{X}}_2$ should reduce these artifacts [43], [44]. The equivariance weight λ specifies the influence of the equivariance term relative to the SURE term.

III. Equivariance2Inverse (E2I)

Equivariance2Inverse (E2I) is a new self-supervised CT reconstruction method that combines ideas from existing methods, with the goals of being accurate, being simple to calibrate, and being robust to sparsity and correlated noise. Its loss consists of a projection-wise cross-validation term, similar to S2I, and an equivariance term, similar to REI.

A. Cross-validation term

A projection-wise cross-validation approach similar to S2I is used, because it does not have parameters that require calibration, while still being robust to sparsity and correlated noise. In every iteration of training of E2I, one angle $\boldsymbol{J}$ will be randomly sampled, and the projection data from this angle $\widetilde{\boldsymbol{Y}}_{\boldsymbol{J}}$ will be used as target data. The projection data of all other angles $\widetilde{\boldsymbol{Y}}_{\boldsymbol{J}^{\text{c}}}$ will be used as input:

$$\mathbb{E}\left[\left\| A_{\boldsymbol{J}}(g(R_{\boldsymbol{J}^{\text{c}}}(\widetilde{\boldsymbol{Y}}_{\boldsymbol{J}^{\text{c}}}))) - \widetilde{\boldsymbol{Y}}_{\boldsymbol{J}} \right\|_2^2\right]. \tag{8}$$

The expected value over $\boldsymbol{J}$ is the average over all possible values of $\boldsymbol{J}$. Therefore, in expectation, this term is the same as the S2I loss with n splits. Like S2I, all data is used as input to the neural network during inference, introducing some bias.

B. Equivariance term

An equivariance term similar to REI (Equation 7) is used to reduce limited-angle and sparse-view artifacts. The equivariance term of E2I is based on different forward model assumptions than REI. It assumes that the pre-processed projection data $\widetilde{\boldsymbol{Y}}$ has additive blurred Gaussian noise as in Equation 4. While this does not follow the forward model in Equation 2 exactly, so it may introduce some bias, the parameters of this forward model are simpler to calibrate (see Sections IV-B2 and VII-A). The resulting equivariance term is $\mathbb{E}\left[\lambda h_{\text{E2I}}\left(\widetilde{\boldsymbol{Y}}_{\boldsymbol{J}^{\text{c}}}, \boldsymbol{J}^{\text{c}}\right)\right]$ with $\lambda h_{\text{E2I}}\left(\widetilde{\boldsymbol{Y}}_{\boldsymbol{J}^{\text{c}}}, \boldsymbol{J}^{\text{c}}\right)$ defined as:

$$\begin{aligned} \widetilde{\boldsymbol{X}}_1 &= t(g(R_{\boldsymbol{J}^{\text{c}}}\widetilde{\boldsymbol{Y}}_{\boldsymbol{J}^{\text{c}}}), \boldsymbol{T}) \\ \widetilde{\boldsymbol{X}}_2 &= g(R(A\widetilde{\boldsymbol{X}}_1 + \boldsymbol{N}_{\text{BG}})) \\ \lambda h_{\text{E2I}}\left(\widetilde{\boldsymbol{Y}}_{\boldsymbol{J}^{\text{c}}}, \boldsymbol{J}^{\text{c}}\right) &= \lambda \left\| \widetilde{\boldsymbol{X}}_1 - \widetilde{\boldsymbol{X}}_2 \right\|_2^2 \end{aligned} \tag{9}$$

The equivariance term is calculated from $\widetilde{\boldsymbol{Y}}_{\boldsymbol{J}^{\text{c}}}$ instead of from $\widetilde{\boldsymbol{Y}}$, so that the result of $g(R_{\boldsymbol{J}^{\text{c}}}\widetilde{\boldsymbol{Y}}_{\boldsymbol{J}^{\text{c}}})$ can be re-used from the calculation of the cross-validation term, making the method more computationally efficient.

The equivariance weight λ specifies the influence of the equivariance term relative to the cross-validation term. If λ is set to a low value, the equivariance term will only have a small effect on the components of $\boldsymbol{X}$ in the range-space of A. However, it can still improve the estimation of the components of $\boldsymbol{X}$ in the null-space of A, because these components do not affect the cross-validation term.

IV. Self-supervised CT benchmark

In this benchmark, the existing self-supervised CT methods were compared with each other, and with E2I, supervised learning (SUP), and an FBP reconstruction.

A. Datasets

1) Synthetic foam datasets: The goal of using these datasets is to test whether the image quality of the methods is negatively affected by sparsity and blurring and limited-angle geometries. Synthetic data was used so that the exact ground truth and the exact forward model parameters were available. Moreover, the model assumptions could be changed one at a time without affecting the further behavior of the model. A limited-angle and a complete geometry were used, with and without blurring, resulting in four combinations. The noise-free projection data and ground truth volume data of a cylinder of foam were generated using the `foam_ct` library [45]. 20 volumes of 256 slices of 256×256 pixels were generated. Two volumes were used for testing, two volumes were used for validation, and the remaining volumes were used for training. 512 projections of width 384 were generated over a range of 180° in a parallel beam geometry. In the complete geometry datasets, all projections were used, and in the limited-angle datasets, the first 256 projections were used, resulting in a 90° missing wedge. The measurement data was generated according to the physics-based forward model in Equation 2, with a constant photon count $\boldsymbol{c}$ of 500, a Gaussian variance $\boldsymbol{v}$ of 50, and $\boldsymbol{u} = 0$ and $\boldsymbol{w} = 1$. In the blurred datasets, B is a convolution with a Gaussian kernel with a standard deviation of 0.8, as was used in [26], and in the blurring-free datasets, B is the identity matrix.

2) Real-world 2DeteCT datasets: The goal of using these datasets is to provide a good indication of how well the methods perform in real-world applications. The 2DeteCT dataset [27] was used, which consists of images of a cardboard tube filled with dried fruits, nuts, and lava stones. The overall shape and contrast of the 2DeteCT data approximate those of a

medical abdominal scan [27], and the individual fruits and nuts have natural variation in shape and texture similar to human organs. Raw 2D fan-beam projection data is available, with every image acquired in three modes: (1) high-noise, (2) low-noise, and (3) no filtering (for testing beam hardening). Data from two of these modes was used as two benchmark datasets. The mode 1 (high-noise) data was used with a complete operator. To limit GPU memory use, the projection images in this dataset were downscaled by a factor of two, and every second projection image was used. The mode 2 (low-noise) data was used with a sparse-view and limited-angle projection operator. This operator used 136 equally spaced projections over a range of 136°, which is similar to a low-cost C-arm acquisition [23].

For both datasets, an FBP reconstruction of all mode 2 data (3600 projections) was used as reference for supervised learning and for calculating error metrics. While this is not a perfect ground truth, it does address the main sources of errors in the 2DeteCT benchmark datasets: The downscaled high-noise benchmark dataset has significantly higher noise, because it uses a 30 times lower tube power than the reference [27]. The limited-angle sparse-view benchmark dataset has roughly 26.5 times fewer projection images than the reference dataset.

During the acquisition of 2DeteCT, the detector of the CT scanner was replaced. Only data from the second detector was used to ensure that the forward model parameters are consistent for all scans. The data from four randomly sampled scanning sessions (200 slices) were used for testing, and four other random sessions were sampled as validation data. The remaining 1770 slices were used for training.

B. Implementation

1) Neural network training: A separate neural network was trained for each combination of method and dataset. The same U-Net architecture [46] was used in all methods, except that the depth and number of channels were selected based on the image resolution of each dataset to limit the GPU memory use. They were chosen so that at the maximum depth, the resolution and number of channels of the layers were roughly the same between the datasets. On the limited-angle 2DeteCT dataset, the network depth was 7, and the number of channels in the first layer was 8. On the complete $2\times$ downscaled 2DeteCT dataset, the network depth was 6, and the number of channels in the first layer was 16. On the foam datasets, the network depth was 4, and the number of channels in the first layer was 64.

The optimizer was the ADAM optimizer [47] with a learning rate of 0.01 and no weight decay. The batch size was 4, which was achieved by parallel training on 4 GPUs (4x Nvidia TITAN X 12GB, 4x Nvidia GTX 1080Ti 11GB, or 4x Nvidia RTX 2080Ti 11GB). Training was stopped after 1000 epochs or when no improvement was observed on the validation loss for 250 epochs. The network weights with the best validation loss were used for inference. PyTorch [48] and PyTorch Lightning [49] were used for the training, and the projection operator A was implemented in a differentiable and matrix-free manner using Tomosipo [50].

2) Forward model parameter calibration: To estimate the blur convolution kernels used in NN2I and E2I, the approach from [24] was used. On the foam data, 1024 images with the same image content but with independent noise were generated for this task. On the 2DeteCT data, a background region with no attenuation of 300 sinograms was used. To estimate the standard deviation of the noise, the projection data was first deconvolved, and then the pixel-wise standard deviation was calculated over the same data.

SURE and REI assumed Poisson + Gaussian noise on raw data. On the synthetic data, the exact gain and Gaussian standard deviation were used. On the 2DeteCT data, the gain was estimated by pixel-wise dividing the variance and mean over 300 sinograms of a background region with no attenuation, and then averaging the pixel-wise results. The Gaussian standard deviation was assumed to be zero.

For E2I and REI, networks were trained with multiple power-of-ten values of the equivariance weight λ on each combination of method and dataset. The network results with the lowest PSNR on the validation set are reported as the benchmark results.

C. Metrics

The mean and standard deviation of the PSNR and the Structural Similarity Index Measure (SSIM) [51] over the images of the test set were calculated as evaluation metrics. The PSNR is inversely related to the supervised learning loss (Equation 5). The SSIM predicts the perceived quality by a human observer.

V. E2I Noise Model Ablation Study

In this ablation study, the contribution of the noise modeling in the equivariance term (Equation 9) was investigated. Two modifications of E2I were tested: In the *E2I (no eq. blur)* modification, non-blurred Gaussian noise was added instead of blurred Gaussian noise in the equivariance term. In the *E2I (no eq. noise)* modification, no noise was added in the equivariance term.

These variants of E2I were compared on the blurred and non-blurred limited-angle foam datasets, and both 2DeteCT datasets. The standard deviation of the non-blurred Gaussian noise was calculated on the same data as the standard deviation of the blurred Gaussian noise in Section IV-B2, but the deconvolution of the projection data was not performed. The same procedures as in the benchmark were used to train and evaluate the neural networks, including optimizing λ over factors of ten based on the validation-set PSNR.

VI. Results

The benchmark results on the synthetic data and the 2DeteCT data are shown in Tables II and III, respectively. Figure 3 shows the PSNR for all tested values of the equivariance weight λ. The ablation study results are shown in Table IV. An example of a reconstruction of each method on each dataset is shown in Figure 4.

TABLE II
BENCHMARK RESULTS ON THE SYNTHETIC DATA.

	Complete		Limited-Angle		Blurred, Complete		Blurred, Limited-Angle	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
FBP	13.72 ± 0.09	0.43 ± 0.01	7.31 ± 0.14	0.19 ± 0.01	16.69 ± 0.11	0.46 ± 0.01	9.65 ± 0.17	0.22 ± 0.01
SUP	29.89 ± 0.36	0.99 ± 0.00	23.36 ± 0.46	0.96 ± 0.00	28.95 ± 0.36	0.99 ± 0.00	22.19 ± 0.48	0.92 ± 0.00
N2I	24.90 ± 0.16	0.86 ± 0.00	9.75 ± 0.18	0.23 ± 0.01	18.72 ± 0.12	0.81 ± 0.00	10.45 ± 0.19	0.23 ± 0.01
S2I	25.48 ± 0.19	0.95 ± 0.00	18.95 ± 0.20	0.72 ± 0.01	**20.46 ± 0.13**	**0.91 ± 0.00**	**17.32 ± 0.24**	**0.67 ± 0.01**
P2P[1]	21.72 ± 0.20	0.91 ± 0.01	17.87 ± 0.45	0.77 ± 0.01	16.78 ± 0.61	0.41 ± 0.01	8.32 ± 0.34	0.13 ± 0.01
NN2I	24.10 ± 0.45	0.74 ± 0.02	17.06 ± 0.36	0.50 ± 0.01	20.27 ± 0.17	0.79 ± 0.02	16.44 ± 0.26	0.51 ± 0.01
SURE[1]	25.74 ± 0.21	**0.96 ± 0.00**	19.36 ± 0.21	0.73 ± 0.01	1.40 ± 0.04	0.12 ± 0.00	-7.82 ± 0.17	0.03 ± 0.00
REI[1,2]	**26.80 ± 0.19**	**0.96 ± 0.00**	**20.32 ± 0.26**	**0.81 ± 0.01**	14.44 ± 0.27	0.60 ± 0.01	12.33 ± 0.27	0.37 ± 0.01
E2I[2]	**26.29 ± 0.22**	0.93 ± 0.00	**22.42 ± 0.38**	**0.92 ± 0.00**	**20.35 ± 0.13**	**0.89 ± 0.00**	**19.18 ± 0.23**	**0.86 ± 0.01**

The best results are shown in underlined boldface, and the second bests in boldface. The methods marked with [1] assume that the noise is pixel-wise independent. The methods marked with [2] use an equivariance loss term, and the table shows the results for the value of λ with the highest PSNR.

TABLE III
BENCHMARK RESULTS ON 2DETECT.

	2× Downscaled, Complete, High-Noise		Limited-Angle, Sparse-View, Low-Noise	
	PSNR	SSIM	PSNR	SSIM
FBP	16.39 ± 0.53	0.05 ± 0.00	17.00 ± 0.49	0.07 ± 0.00
SUP	33.67 ± 0.69	0.78 ± 0.01	30.37 ± 0.63	0.59 ± 0.02
N2I	**33.66 ± 0.69**	**0.78 ± 0.01**	23.77 ± 1.13	0.30 ± 0.04
S2I	28.60 ± 1.21	0.64 ± 0.03	28.05 ± 0.75	0.46 ± 0.02
P2P[1]	17.55 ± 0.34	0.08 ± 0.00	17.10 ± 0.35	0.06 ± 0.01
NN2I	17.71 ± 1.41	0.08 ± 0.02	28.03 ± 0.81	**0.49 ± 0.02**
SURE[1]	5.45 ± 1.43	0.00 ± 0.00	21.09 ± 0.39	0.09 ± 0.01
REI[1,2]	28.56 ± 0.62	0.58 ± 0.01	**28.50 ± 0.81**	0.48 ± 0.02
E2I[2]	**32.60 ± 0.67**	**0.69 ± 0.02**	**29.21 ± 0.71**	**0.51 ± 0.02**

The best results are shown in underlined boldface, and the second bests in boldface. The methods marked with [1] assume that the noise is pixel-wise independent. The methods marked with [2] use an equivariance loss term, and the table shows the results for the value of λ with the highest PSNR.

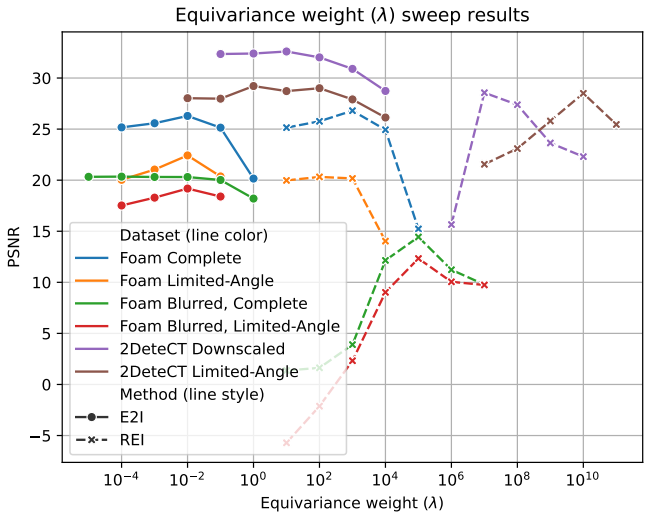

Fig. 3. The test-set PSNR of neural networks trained with different values of the equivariance weight λ. A different line color was used for each dataset, and a different line style for each method.

A. Blurring and noise model assumptions

S2I, P2P, NN2I, and SURE calculate the loss in the projection domain, but make different noise model assumptions. We will compare these four methods to show the effects of these noise model assumptions. The synthetic data without blurring (left two columns of Table II) matches the noise model assumptions of SURE exactly; in this case, SURE should be an unbiased estimator, which explains why it performs best among these four methods. When scintillator blurring is simulated (right two columns of Table II), the noise is no longer pixel-wise independent, so the methods that assumed pixel-wise independence perform worse. Of the four methods, S2I achieves the second-highest PSNR on the synthetic datasets without blurring and the highest PSNR on the synthetic datasets with blurring. On the 2DeteCT data, blurring is present (Figure 1), which could explain why SURE had a lower PSNR than S2I, and NN2I. However, it does not explain the low performance of SURE on the Complete, high-noise, downscaled data (Table III), because the downscaling reduces the effects of blurring. Other causes may be calibration errors or unmodeled effects, such as scattering. S2I again showed the highest PSNR of these four methods (Table III).

B. Limited-angle data and the equivariance loss term

S2I is similar to N2I, but it should be less affected by A having a non-trivial null space [15]. This explains why the difference between the performance (PSNR and SSIM) of N2I and S2I is much bigger on the synthetic limited-angle data than on the synthetic complete data (Table II, both with and without blurring). The equivariance term of REI was designed to make SURE more robust to A having a non-trivial null space [13]. The effect of the equivariance term can be seen on the limited-angle 2DeteCT reconstructions (bottom row of Figure 4), where only the self-supervised methods with an equivariance term (REI and E2I) correctly reconstructed the tube as a continuous circle. In all our experiments, REI outperformed SURE, so the equivariance term may also be beneficial for complete geometries. On the blurred synthetic data, it appears that the equivariance term compensates to some degree for the incorrect noise-model assumptions of the SURE loss term, because on that data the PSNR and SSIM of REI are much higher than those of SURE (Table II), and the resulting images are a lot less noisy (Figure 4). Nevertheless,

TABLE IV
E2I NOISE MODEL ABLATION STUDY RESULTS.

	Foam Limited-Angle		Foam Blurred, Limited-Angle		2DeteCT 2x Downscaled, Complete, High-Noise		2DeteCT Limited-Angle, Sparse-View, Low-Noise	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
E2I	22.42 ± 0.38	0.92 ± 0.00	19.18 ± 0.23	0.86 ± 0.01	32.60 ± 0.67	0.69 ± 0.02	29.21 ± 0.71	0.51 ± 0.02
E2I (no eq. blur)	N.A.	N.A.	19.04 ± 0.23	0.86 ± 0.01	32.58 ± 0.68	0.67 ± 0.02	29.36 ± 0.69	0.51 ± 0.02
E2I (no eq. noise)	21.84 ± 0.35	0.92 ± 0.00	18.09 ± 0.25	0.76 ± 0.01	32.58 ± 0.68	0.68 ± 0.02	29.34 ± 0.70	0.51 ± 0.03

The E2I results are the same as the corresponding results in Tables II and III.

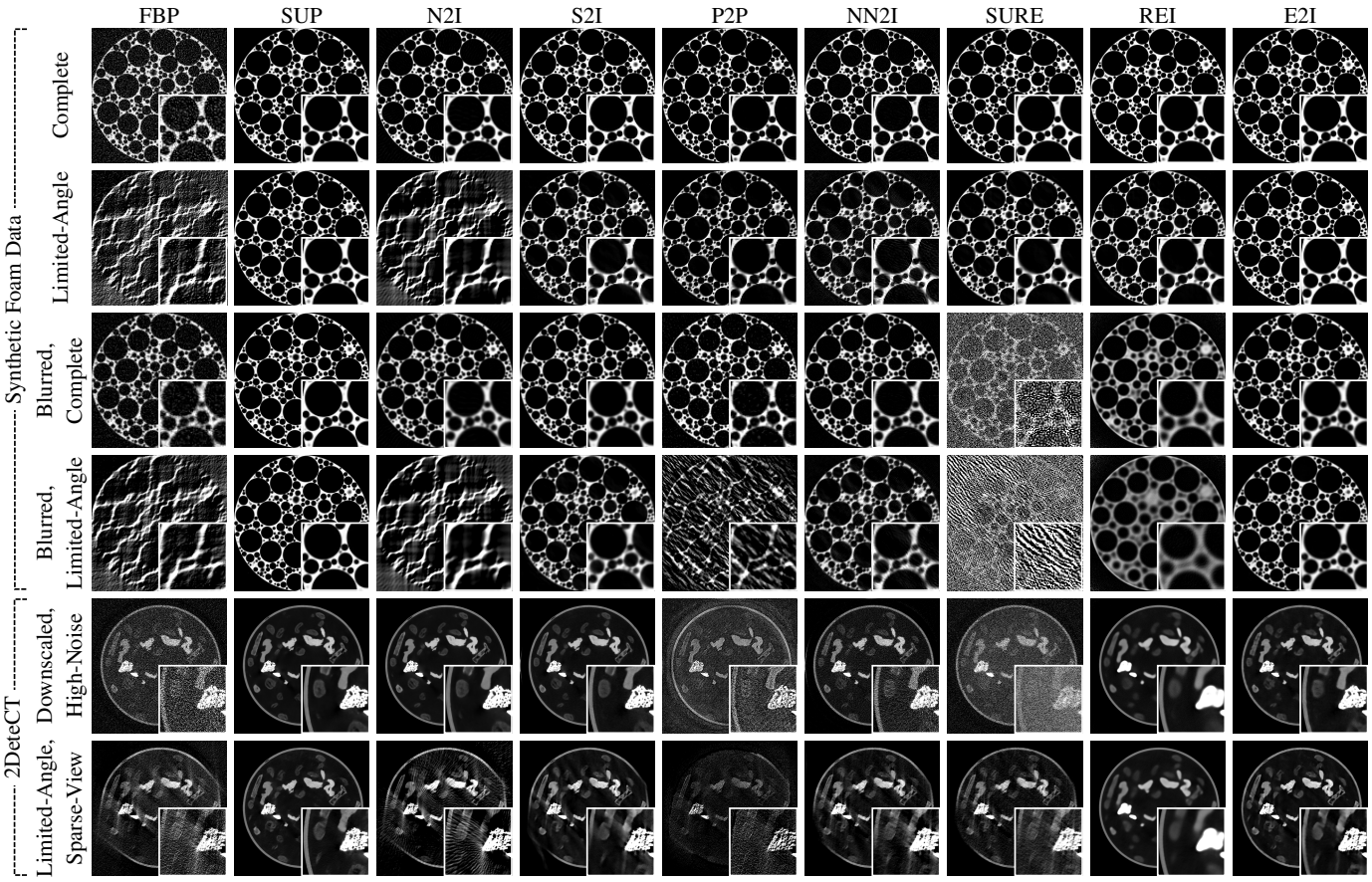

Fig. 4. A reconstruction from each method on the first image of the test set of each dataset. The insets provide a two times magnified view of the center-left side of the objects. The inset on the 2DeteCT data shows a lava stone (white), some dried fruits or nuts (grey), and the edge of the cardboard tube (near the left edge), which in the limited-angle reconstructions may be incomplete because of limited-angle artifacts.

multiple other methods had better results on the same data. The E2I loss consists of a loss term similar to S2I and an equivariance term. On the Blurred, Complete synthetic data, the performance of S2I was slightly higher than that of E2I, but on all other datasets in our benchmark, E2I performed better.

C. The effect of the equivariance weight

Figure 3 shows the test-set PSNR of the networks for the REI and E2I methods that were trained with different values of λ. The value of λ with the highest PSNR was the same on the validation set as on the test set in all cases, so the highest point of every curve in Figure 3 corresponds to the benchmark results in Tables II and III. The fact that the optimal value of λ was generally higher for REI than for E2I can be explained by the fact that the data consistency term of REI is calculated using the raw data, while the data consistency loss of E2I is calculated on pre-processed data, which typically has lower values. When comparing the optimal value of λ with the adjacent power of ten values, the decrease in PSNR is generally larger for REI than for E2I, showing that E2I is less sensitive to tuning λ.

D. E2I noise model ablation study

The results in Table IV show that on the foam datasets, adding noise in the equivariance term improves the PSNR and SSIM, but the difference between adding blurred or unblurred noise is very small. On the 2DeteCT datasets, the PSNR and SSIM of the three E2I variants are very similar. These results suggest that the distribution of the added noise in the E2I equivariance term affects the results only slightly.

E. Computational costs

Table V shows the number of calls to the neural network g. Additionally, it shows the computation time per iteration and the GPU memory use during training on the synthetic Foam, Blurred, Complete dataset. The N2I, S2I, P2P, and NN2I methods all use the same number of neural network calls and require a similar amount of GPU memory. The use of an equivariance term in the loss adds one additional neural network call, which increases the computation time and GPU memory use. SURE is calculated using a Monte Carlo-based estimate of the divergence term [13], which requires three additional calls to the neural network, strongly increasing the computation time and GPU memory use. When summing up the time used on different computers, the total training time of all neural networks in this paper is approximately eleven months.

TABLE V
THE COMPUTATIONAL COSTS OF TRAINING THE BENCHMARKED METHODS.

Method	**NN Calls**	**Time per Iteration** (ms)	**GPU Memory Use** (MiB)
N2I	1	87.4	912
S2I	1	88.0	912
P2P	1	121.6	916
NN2I	1	72.5	918
SURE	4	236.8	1828
REI	5	292.2	2144
E2I	2	132.4	1233

The GPU memory use and computation time per iteration are measured for a batch size of one per GPU on four Nvidia Titan X GPUs on the blurred and complete synthetic foam dataset.

VII. DISCUSSION

A. The importance of calibration

An inherent difficulty of applying a method to a new dataset is finding the best parameter values for running the method. Moreover, none of the methods specified a calibration approach for their model parameters. On the synthetic foam data, the data generation parameters were used for SURE and REI, and extensive additional calibration measurements were generated to estimate the parameters for NN2I and E2I. Therefore, we expect that the results on the generated data are not strongly affected by calibration errors. On the 2DeteCT data, no exact parameters or calibration measurements were available. Therefore, calibration inaccuracies may have had a larger impact on these results.

REI and SURE depend on the parameters of a Poisson + Gaussian noise model that would have required many additional calibration measurements to estimate accurately [26], [52]. The Poisson component of X-ray detector noise is typically much larger than the Gaussian component [53], which led us to configure these methods with $\sigma = 0$. The calibration for the parameters of E2I and NN2I on 2DeteCT was done using a background region with no attenuation, so no additional calibration measurements were required. The recently presented UNSURE method [20] proposes to optimize the model parameters of SURE-type optimizers alongside optimizing the neural network, removing the need for calibration.

B. Extending the forward model

There is currently no consensus among self-supervised CT reconstruction methods on what forward model assumptions to make. This raises the question whether more aspects of X-ray physics should be modeled.

Beam hardening [33] is a common artifact, so it would be interesting future work to study how self-supervised CT reconstruction methods are affected by it, and if it could be corrected by self-supervised learning. The 2DeteCT dataset contains the mode 3 data that was acquired specifically for benchmarking beam hardening reduction [27].

Scintillator blurring could be modeled in more detail by taking into account that it is slightly angle dependent [54]. Moreover, the NN2I and E2I methods take into account that scintillator blurring results in correlated noise, but they do not account for the fact that the signal component ($\exp(-A\boldsymbol{x})$ in Equation 2) is also blurred, which led to slightly blurry reconstructions on the synthetic blurred data (Figure 4).

The focal spot of the X-ray source [55] and scattering [25] may also cause blurring. However, both of these effects only cause blurring of the signal component and not of the noise, so they can not explain the correlated background noise in Figure 1. Scattering is also material dependent [25], and the radius of the blur is much larger than that of scintillator blurring [25], [56], [57].

The modeling of this paper assumes that an energy-integrating detector is used, which is currently the most widely used type of X-ray detector in CT scanners. Photon-counting detectors, which are increasingly being used in CT scanners [58], [59], do not have a scintillator and therefore do not exhibit scintillator blurring. Photon-counting detectors might exhibit other effects that could require separate modeling, but that is considered outside of the scope of this paper.

C. Estimation using equivariance

The main aim of the equivariance terms in REI and E2I is to make the network learn how to estimate the components of $\boldsymbol{X}$ that are in the null-space of A, which should reduce blurring and streaking artifacts in limited-angle or sparse-view reconstruction. Necessary and sufficient conditions on the operator A and the invariance of the distribution of $\boldsymbol{X}$ for learning to estimate the null-space components using self-supervised learning were formulated in [44]. No guarantees about the bias or statistical consistency of REI-like equivariance terms have been derived, but they have been used successfully in several inverse problems [13], [43], [60], [61].

The recently introduced Equivariant Splitting (ES) self-supervised reconstruction method [17] has been proven to converge to the supervised learning loss of linear inverse problems with an invariant data distribution even when the projection operator A has a non-trivial null-space. It works by using the invariance to interpret data from one operator as being collected from multiple operators [62]. It would be interesting future work to combine aspects of ES and E2I into a method that is both robust to scintillator blurring and an unbiased estimator of the null-space components.

VIII. Conclusion

The benchmark in this paper evaluated recent self-supervised CT reconstruction methods on synthetic data with and without scintillator blurring and a limited-angle geometry, and on two real-world datasets from 2DeteCT. REI, which is SURE with an additional equivariance term, had a better performance (PSNR and SSIM) than SURE on all benchmark datasets (Tables II & III). SURE makes strong model assumptions (pixel-wise independent Poisson + Gaussian noise with known parameters), and it was the best-performing method without an equivariance term on the non-blurred synthetic data, where these assumptions were met exactly. However, on the other datasets, where these assumptions were not met exactly, SURE was outperformed by multiple other methods. S2I, on the other hand, had the most general model assumptions (projection-wise independent zero-mean noise), and it performed best or second best of the methods without an equivariance term on all benchmark datasets. The E2I method introduced in this paper combines the robustness of S2I with the performance increase of the equivariance term of REI. The PSNR of E2I was the best or a close second-best (at most 1.06 lower) on all benchmark datasets.

IX. Code and data availability

The code is available on Github at: https://github.com/D1rk123/equivariance2inverse.
The synthetic foam data is available on Zenodo [63] at: https://zenodo.org/records/16735632. The 2DeteCT dataset [27] is available on Zenodo at: https://zenodo.org/records/8014758.

Acknowledgments

We gratefully acknowledge our colleagues: Allard Hendriksen for introducing us to the topic of self-supervised CT reconstruction, and Marcos Obando for the useful discussions and feedback.

References

[1] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE transactions on image processing*, vol. 26, no. 9, pp. 4509–4522, 2017.

[2] J. Adler and O. Öktem, "Learned primal-dual reconstruction," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1322–1332, 2018.

[3] G. Ongie, A. Jalal, C. A. Metzler, R. G. Baraniuk, A. G. Dimakis, and R. Willett, "Deep learning techniques for inverse problems in imaging," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 39–56, 2020.

[4] Y. Song, L. Shen, L. Xing, and S. Ermon, "Solving inverse problems in medical imaging with score-based generative models," in *International Conference on Learning Representations*, 2022.

[5] H. Chung, B. Sim, D. Ryu, and J. C. Ye, "Improving diffusion models for inverse problems using manifold constraints," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 683–25 696, 2022.

[6] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye, "Diffusion posterior sampling for general noisy inverse problems," *International Conference on Learning Representations*, 2023.

[7] J. Song, A. Vahdat, M. Mardani, and J. Kautz, "Pseudoinverse-guided diffusion models for inverse problems," in *International Conference on Learning Representations*, 2023.

[8] Y. Sun, J. Liu, M. Xie, B. Wohlberg, and U. S. Kamilov, "Coil: Coordinate-based internal learning for tomographic imaging," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 1400–1412, 2021.

[9] G. Zang, R. Idoughi, R. Li, P. Wonka, and W. Heidrich, "Intratomo: self-supervised learning-based tomography via sinogram synthesis and prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1960–1970.

[10] R. Zha, Y. Zhang, and H. Li, "Naf: neural attenuation fields for sparse-view cbct reconstruction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 442–452.

[11] Q. Wu, R. Feng, H. Wei, J. Yu, and Y. Zhang, "Self-supervised coordinate projection network for sparse-view computed tomography," *IEEE Transactions on Computational Imaging*, vol. 9, pp. 517–529, 2023.

[12] A. A. Hendriksen, D. M. Pelt, and K. J. Batenburg, "Noise2inverse: Self-supervised deep convolutional denoising for tomography," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1320–1335, 2020.

[13] D. Chen, J. Tachella, and M. E. Davies, "Robust equivariant imaging: a fully unsupervised framework for learning to image from noisy and partial measurements," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5647–5656.

[14] M. O. Unal, M. Ertas, and I. Yildirim, "Proj2proj: self-supervised low-dose ct reconstruction," *PeerJ Computer Science*, vol. 10, p. e1849, 2024.

[15] N. Gruber, J. Schwab, E. Gizewski, and M. Haltmeier, "Sparse2inverse: Self-supervised inversion of sparse-view ct data," 2024. [Online]. Available: https://arxiv.org/abs/2402.16921

[16] N. Gruber, J. Schwab, M. Haltmeier, A. Biguri, C. Dlaska, and G. Hwang, "Noisier2inverse: Self-supervised learning for image reconstruction with correlated noise," *IEEE Access*, 2025.

[17] V. Sechaud, J. Scanvic, Q. Barthélemy, P. Abry, and J. Tachella, "Equivariant splitting: Self-supervised learning from incomplete data," in *The Fourteenth International Conference on Learning Representations*, 2026.

[18] H. Xu and A. Perelli, "Rotational augmented noise2inverse for low-dose computed tomography reconstruction," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 8, no. 2, pp. 208–221, 2023.

[19] P. C. Hansen, J. Jørgensen, and W. R. Lionheart, *Computed Tomography: Algorithms, Insight, and Just Enough Theory*. SIAM, 2021.

[20] J. Tachella, M. Davies, and L. Jacques, "Unsure: self-supervised learning with unknown noise level and stein's unbiased risk estimate," in *The Thirteenth International Conference on Learning Representations*, 2025.

[21] A. Wirgin, "The inverse crime," 2004. [Online]. Available: https://arxiv.org/abs/math-ph/0401050

[22] J. Nuyts, B. De Man, J. A. Fessler, W. Zbijewski, and F. J. Beekman, "Modelling the physics in the iterative reconstruction for transmission computed tomography," *Physics in Medicine & Biology*, vol. 58, no. 12, p. R63, 2013.

[23] M. Abella, C. de Molina, N. Ballesteros, A. García-Santos, Á. Martínez, I. Garcia, and M. Desco, "Enabling tomography with low-cost c-arm systems," *PLoS One*, vol. 13, no. 9, p. e0203817, 2018.

[24] A. Graas and F. Lucka, "Scintillator decorrelation for self-supervised x-ray radiograph denoising," *Measurement Science and Technology*, 2025.

[25] V. Andriiashen, R. van Liere, T. van Leeuwen, and K. J. Batenburg, "Quantifying the effect of x-ray scattering for data generation in real-time defect detection," *Journal of X-ray Science and Technology*, vol. 32, no. 4, pp. 1099–1119, 2024.

[26] ——, "X-ray image generation as a method of performance prediction for real-time inspection: a case study," *Journal of Nondestructive Evaluation*, vol. 43, no. 3, p. 79, 2024.

[27] M. B. Kiss, S. B. Coban, K. J. Batenburg, T. van Leeuwen, and F. Lucka, "2detect-a large 2d expandable, trainable, experimental computed tomography dataset for machine learning," *Scientific data*, vol. 10, no. 1, p. 576, 2023.

[28] G. Poludniowski, A. Omar, R. Bujila, and P. Andreo, "Spekpy v2. 0—a software toolkit for modeling x-ray tube spectra," *Medical Physics*, vol. 48, no. 7, pp. 3630–3637, 2021.

[29] T. Gomi, K. Koshida, T. Miyati, J. Miyagawa, and H. Hirano, "An experimental comparison of flat-panel detector performance for direct and indirect systems (initial experiences and physical evaluation)," *Journal of Digital Imaging*, vol. 19, pp. 362–370, 2006.

[30] A. Howansky, A. Lubinsky, K. Suzuki, S. Ghose, and W. Zhao, "An apparatus and method for directly measuring the depth-dependent gain and spatial resolution of turbid scintillators," *Medical physics*, vol. 45, no. 11, pp. 4927–4941, 2018.

[31] European Machine Vision Association, "Emva standard 1288: Standard for characterization of image sensors and cameras," 2021, release 4.0 (Linear).

[32] J.-P. Moy and B. Bosset, "How does real offset and gain correction affect the dqe in images from x-ray flat detectors?" in *Medical Imaging 1999: Physics of Medical Imaging*, vol. 3659. SPIE, 1999, pp. 90–97.

[33] T. M. Buzug, *Computed tomography: from photon statistics to modern cone-beam CT*. Springer, 2008.

[34] H. K. Tuy, "An inversion formula for cone-beam reconstruction," *SIAM Journal on Applied Mathematics*, vol. 43, no. 3, pp. 546–552, 1983.

[35] B. D. Smith, "Image reconstruction from cone-beam projections: necessary and sufficient conditions and reconstruction methods," *IEEE transactions on medical imaging*, vol. 4, no. 1, pp. 14–25, 1985.

[36] R. A. Crowther, D. DeRosier, and A. Klug, "The reconstruction of a three-dimensional structure from projections and its application to electron microscopy," *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, vol. 317, no. 1530, pp. 319–340, 1970.

[37] J. Batson and L. Royer, "Noise2self: Blind denoising by self-supervision," in *International Conference on Machine Learning*. PMLR, 2019, pp. 524–533.

[38] A. Krull, T.-O. Buchholz, and F. Jug, "Noise2void-learning denoising from single noisy images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2129–2137.

[39] A. A. Hendriksen, M. Bührer, L. Leone, M. Merlini, N. Vigano, D. M. Pelt, F. Marone, M. Di Michiel, and K. J. Batenburg, "Deep denoising for multi-dimensional synchrotron x-ray tomography without high-quality reference data," *Scientific reports*, vol. 11, no. 1, p. 11895, 2021.

[40] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *The annals of Statistics*, pp. 1135–1151, 1981.

[41] B. Monroy, J. Bacca, and J. Tachella, "Generalized recorrupted-to-recorrupted: Self-supervised learning beyond gaussian noise," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 28 155–28 164.

[42] Y. Le Montagner, E. D. Angelini, and J.-C. Olivo-Marin, "An unbiased risk estimator for image denoising in the presence of mixed poisson–gaussian noise," *IEEE Transactions on Image processing*, vol. 23, no. 3, pp. 1255–1268, 2014.

[43] D. Chen, J. Tachella, and M. E. Davies, "Equivariant imaging: Learning beyond the range space," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4379–4388.

[44] J. Tachella, D. Chen, and M. Davies, "Sensing theorems for unsupervised learning in linear inverse problems," *Journal of Machine Learning Research*, vol. 24, no. 39, pp. 1–45, 2023.

[45] D. M. Pelt, A. A. Hendriksen, and K. J. Batenburg, "Foam-like phantoms for comparing tomography algorithms," *Synchrotron Radiation*, vol. 29, no. 1, pp. 254–265, 2022.

[46] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. [Online]. Available: https://arxiv.org/abs/1412.6980

[48] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[49] W. Falcon and T. P. L. team, "Pytorch lightning," 2024. [Online]. Available: https://doi.org/10.5281/zenodo.10779019

[50] A. A. Hendriksen, D. Schut, W. J. Palenstijn, N. Viganó, J. Kim, D. M. Pelt, T. Van Leeuwen, and K. Joost Batenburg, "Tomosipo: fast, flexible, and convenient 3d tomography for complex scanning geometries in python," *Optics Express*, vol. 29, no. 24, pp. 40 494–40 513, 2021.

[51] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[52] A. Jezierska, C. Chaux, J.-C. Pesquet, and H. Talbot, "An em approach for poisson-gaussian noise modeling," in *2011 19th European Signal Processing Conference*. IEEE, 2011, pp. 2244–2248.

[53] Q. Ding, Y. Long, X. Zhang, and J. A. Fessler, "Modeling mixed poisson-gaussian noise in statistical image reconstruction for x-ray ct," *Arbor*, vol. 1001, no. 48109, p. 6, 2016.

[54] M. Freed, S. Park, and A. Badano, "A fast, angle-dependent, analytical model of csi detector response for optimization of 3d x-ray breast imaging systems," *Medical physics*, vol. 37, no. 6Part1, pp. 2593–2605, 2010.

[55] K. A. Mohan, R. M. Panas, and J. A. Cuadra, "Saber: a systems approach to blur estimation and reduction in x-ray imaging," *IEEE Transactions on Image Processing*, vol. 29, pp. 7751–7764, 2020.

[56] J. A. Seibert and J. M. Boone, "X-ray imaging physics for nuclear medicine technologists. part 2: X-ray interactions and image formation," *Journal of nuclear medicine technology*, vol. 33, no. 1, pp. 3–18, 2005.

[57] A. Malusek, "Calculation of scatter in cone beam ct: Steps towards a virtual tomograph," Ph.D. dissertation, Linköping University, 2008.

[58] J. Greffier, A. Viry, A. Robert, M. Khorsi, and S. Si-Mohamed, "Photon-counting ct systems: a technical review of current clinical possibilities," *Diagnostic and interventional imaging*, vol. 106, no. 2, pp. 53–59, 2025.

[59] E. Wehrse, L. Klein, L. Rotkopf, W. Wagner, M. Uhrig, C. Heußel, C. Ziener, S. Delorme, S. Heinze, M. Kachelrieß *et al.*, "Photon-counting detectors in computed tomography: from quantum physics to clinical practice," *Der Radiologe*, vol. 61, no. Suppl 1, pp. 1–10, 2021.

[60] V. Sechaud, L. Jacques, P. Abry, and J. Tachella, "Equivariance-based self-supervised learning for audio signal recovery from clipped measurements," in *2024 32nd European Signal Processing Conference (EUSIPCO)*. IEEE, 2024, pp. 852–856.

[61] A. Wang and M. Davies, "Perspective-equivariance for unsupervised imaging with camera geometry," in *European Conference on Computer Vision*. Springer, 2024, pp. 116–134.

[62] C. Millard and M. Chiew, "A theoretical framework for self-supervised mr image reconstruction using sub-sampling via variable density noisier2noise," *IEEE transactions on computational imaging*, vol. 9, pp. 707–720, 2023.

[63] D. E. Schut, "Synthetically generated foam ct dataset," 2025. [Online]. Available: https://doi.org/10.5281/zenodo.16735632

Dirk Elias Schut is a PhD researcher at the Computational Imaging group of the Centrum Wiskunde & Informatica (CWI) in the Netherlands. His research is part of the UTOPIA project on bringing per product CT imaging for quality control to food processing factories. He received a Bsc. in Computer Science and a double Msc. in Computer Science and Electrical Engineering, both from Delft University of Technology.

Adriaan Graas received an MSc. Mathematics from Utrecht University, and is a Ph.D. researcher in the Computational Imaging group at CWI, in Amsterdam. His research, on the topic of Mathematics and Algorithms for 3D Imaging of Dynamic Processes, is conducted within the NDNS+ cluster of mathematics research in the Netherlands.

Robert van Liere received the PhD degree in computer science from the University of Amsterdam. He is a principal investigator at the Computational Imaging group of the Centrum Wiskunde & Informatica (CWI) and emeritus full professor at the Eindhoven University of Technology. His research interests include interactive visualization, virtual environments, and human-computer interaction.

Tristan van Leeuwen is the group leader of the Computational Imaging group at the Centrum Wiskunde & Informatica (CWI) and a full professor at Utrecht University. He received his BSc. and MSc. in Computational Science from Utrecht University. He obtained his PhD. in geophysics at Delft University in 2010 and was a postdoctoral researcher at the University of British Columbia in Vancouver, Canada and the CWI. His research interests include: inverse problems, computational imaging, tomography and numerical optimization.