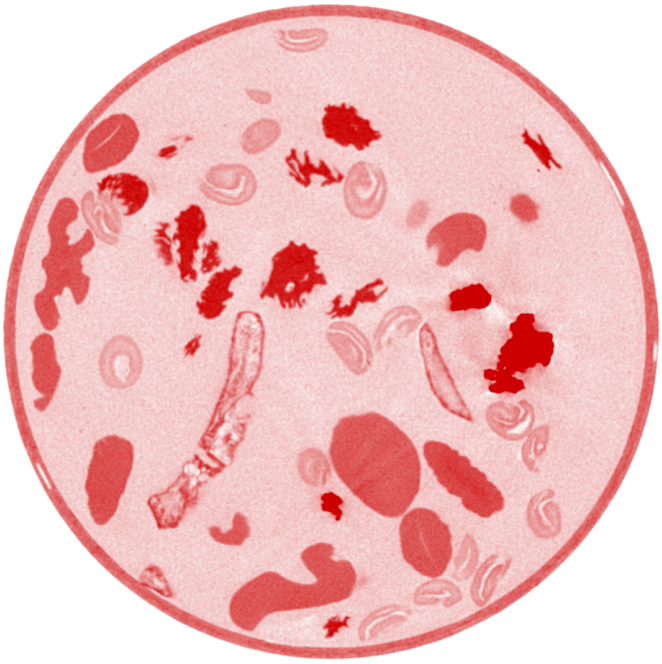
# Advancing learned algorithms for 2D X-ray computed tomography

Kiss, M.B.



# 5
# Learned denoising with simulated and experimental low-dose CT data

Computed tomography has proven itself as a powerful non-invasive imaging technique in many fields such as materials science, industrial testing, and medicine. It uses X-ray technology to create detailed cross-sectional images of the scanned object using computational methods. Since it uses harmful radiation the imposed dose on objects and patients raises concerns and safety guidelines have been established to minimize radiation exposure [32, 52]. The ALARA principle [140], which stands for "As Low As Reasonably Achievable" advises healthcare providers to use the lowest possible radiation dose necessary to produce high-quality images. However, the minimization of radiation dose through lowering the tube current or exposure time seriously degrades the resulting CT images if no corresponding noise compensation is applied before or during image reconstruction [82, 127]. Noisy images can also occur when there are, for example, constraints on the available time or the number of projection angles. In either setting, it is desirable to reduce the amount of noise through computational methods.

Like in many other research fields, recent developments in computational imaging have focused on developing machine learning (ML) approaches to tackle its main challenges. To improve the performance of algorithms, ML methods are used for different image

This chapter is based on:
M. B. Kiss, A. Biguri, C.-B. Schönlieb, K. J. Batenburg, and F. Lucka. "Learned denoising with simulated and experimental low-dose CT data". *arXiv preprint arXiv:2408.08115* (2024).

processing tasks. These tasks are for example segmentation, artifact removal, or noise reduction.

Generally, these ML methods heavily rely on the availability of high-quality data on which they are trained. When there is a lack of such data, usually existing data is augmented, or new data is generated artificially through simulations. These simulations mimic the problem the ML algorithms shall solve and try to resemble real-world data as good as possible.

The fundamental question arising from this approach is to which extent algorithms trained on simulated data are applicable to real-world experimental data. This work is investigating the performance of noise reduction for two common convolutional neural networks (CNNs). These networks are trained on either simulated or experimental noisy data and are applied to both experimental and simulated noisy data.

Typically, researchers would not have access to raw measurement data because CT manufacturers consider them proprietary. This severely limited both, the analysis of noise simulations but also the performance comparison of algorithms trained on simulated data [207]. The data used in this work are 2D slices of X-ray computed tomography images published in the carefully designed study "2DeteCT – A large 2D experimental, trainable and expandable CT data collection for machine learning" [98]. This experimental data was acquired by the group for Computational Imaging at the Centrum Wiskunde & Informatica and is openly available on zenodo [102–113]. The data collection consists of 5,000 distinct image slices acquired in three different modes. The resulting images are either clean, noisy, or artifact-inflicted.

Using the paired data of clean and noisy images, we create a setting for supervised learning that the CNNs can be trained on for noise reduction. In this work, the clean data is used as a measurement basis to add computationally fast, yet accurate simulated noise. With this data collection and the newly simulated data we have three types of sinograms available: an experimental noisy, an experimental clean and a simulated noisy sinogram. We can pair those sinograms with the clean reconstructed target images to show the difference between training on simulated noisy data and using experimental noisy data for the task of learned denoising.

In this paper, we utilize a large 2D computed tomography dataset for machine learning to carry out for the first time a comprehensive study on the differences between the observed performances of algorithms trained on simulated noisy data and on experimental noisy data. For this we train two common neural networks such as the generic U-Net [173] and the more tailored MSD-Net [160] on both types of noisy data, experimental and simulated. These networks are applied to the data they have been trained on but also to their respective counterparts. The evaluation follows via quantitative metrics in the sinogram and reconstructed image domain as well as qualitative visual inspection in the reconstructed image domain only.

The structure of this study is as follows: After a brief overview of related work in noise modelling and mitigation in the field of computed tomography, we focus on the pre-processing of computed tomography data, previous noise simulation approaches

and how they influenced our choices for simulating noisy training data. In the following subsections we describe the preparation of our training data, the method development, the employed comparison metrics, and how we set up the computational experiments. In the results and discussion section, we present the empirical selection of the noise level for our simulated noisy data and analyse the performance of the differently trained networks applied to the two types of noisy data. We focus on three aspects in our analysis: The choice for the evaluation domain, the influence of the image content, the choice of the training setting.

## 5.1 Methods



### 5.1.1 Related work

There is a vast amount of literature investigating the theoretical derivation of accurate noise models for computed tomography images. Generally, they agree that the image noise is directly related to the imaging process and its design criteria such as exposure time, pixels size, slice width, and reconstruction algorithm [58]. Faulkner et al. [58] therefore distinguish between algorithmic and non-algorithmic contributions to noise, and between spatial as well as statistical errors in a CT scan. They note that the statistical noise in the reconstructed images is independent of the number of projections and that the uncertainty is only dependent on the total number of detected photons. Hsieh [83] distinguishes between two principal sources of noise in CT measurement data: quantum noise and electronic noise. Yu et al. [221] showed that the latter usually can be neglected except when the number of detected photons is low and approaches the electronic noise floor. Furthermore, they emphasize that the major difficulty in simulating very-low dose CT measurement data is photon starvation artifacts. These become apparent in reconstructed image slices as ripples or rings in the central region or streaking artifacts between high-density regions. Yu et al. [221] furthermore concluded that their proposed method is not able to simulate images with very-low dose because the photon starvation artifacts are quite complicated. Additionally, they reiterated the call of numerous researchers to get access to raw CT data to allow for testing algorithms for iterative reconstruction and noise reduction [155]. Manufacturers of clinical CT scanners usually introduce nonlinear filters [83] on the measured data to counter beam-hardening and photon starvation artifacts. Therefore, real-world experimental raw data prior to this nonlinear filtering would enable more accurate noise simulations but are usually unavailable.

In practice, it is very challenging to bound the concept of noise in CT image reconstruction from artifacts originating from sources such as sample movement, geometric misalignment, or under-sampling. In this work, we choose to confine our investigations purely to the noise in the sinograms induced by the photon detection in the detector. We note, however, that in reality it is hard to completely disentangle these artifacts and their origins.

### 5.1.2 Noise simulation

Zainulina et al. [223] concluded in their work that adding noise to the images artificially could bias the predictions of a convolutional neural network (CNN) depending on the accuracy of the noise simulation. This noise simulation requires an in-depth understanding of the actual CT system and might not be feasible at times. The noise in low-dose CT measurement data is influenced by many factors such as the quantum noise, the logarithmic transformation of the measurements, or pre-reconstruction corrections for system calibration which makes modeling the noise in the reconstructed images particularly challenging [130].



**Pre-processing** A common practice to pre-process the projection data, consisting of raw photon counts per detector pixel, is the so-called dark- and flat-field correction. The dark-fields (D in Equation 5.1) represent the offset counts of the detector system and the flat-fields (F in Equation 5.1) are the values measured when irradiating the detector without an object present between the X-ray source and the detector. These two additional measurements are usually acquired before and/or after the acquisition of the 360°projections and used to remove the dark currents of the detector and to normalize its pixel-dependent sensitivities. With this, the sinograms (S) can then be converted into a beam intensity loss image (ILI) following the Beer-Lambert law after applying the negative logarithm to it according to the formula:

$$y = -\log\left(ILI\right) = -\log\left(\frac{S-D}{F-D}\right) \tag{5.1}$$

Such calibrated projection data no longer follows a compound Poisson distribution but is close to a Gaussian distribution with signal-dependent variance [127]. Furthermore, it has been shown that particularly the logarithm operation significantly amplifies the noise when the signal is low [82]. If we want to denoise the low-dose CT measurements before reconstruction, this is best done in the stage of the beam intensity loss image (ILI), so before taking the negative logarithm. If we would denoise the X-ray absorption sinogram ($y$) instead of the beam intensity loss image (ILI), the application of the negative logarithm would have amplified the noise and changed its distribution.

Based on these findings and considerations we pre-process the raw sinograms of the experimental noisy measurements to beam intensity loss images (ILI) as shown in Equation 5.1 and apply the denoising before taking the negative logarithm. For the preparation of the simulated noisy sinograms we pre-process the sinograms of the high-dose CT measurements of "mode 2" with dark- and flat-field corrections before applying simulated noise to them. With this we prevent our data to experience distributional shifts that might influence the performance of the denoising networks.

**Noise simulation approaches** To date, there have been proposed several different approaches to simulate the noise in CT measurement data [7, 17, 63, 121, 137, 138, 140, 206–208]. Under the condition that the raw data of a high-dose and low-noise scan is available many studies simulated low-dose and high-noise projection data by

applying synthetic Poisson noise or a combination of synthetic Poisson and Gaussian noise to a high-dose scan [53, 54]. The common models for noise simulation use a relatively simple model of CT acquisition considering a monochromatic X-ray source. This source generates photons that are attenuated by a scanned object and detectors counting surviving photons which are governed by Poisson statistics. More complicated methods range from a detailed characterization of signal statistics of X-ray CT [206–208] over noise equivalent quanta [137, 138] to accounting for energy-integrating detectors [121, 137]. The interested reader may be pointed to the study of Zabic et al. giving a broad overview on the state-of-the-art [222].

To motivate our noise simulation approach we highlight what approaches have been used in practice by previous publications in the field. In particular, there are three noise challenges that have been conducted in the past ten years that have attracted attention to deep learning based denoising. Firstly, the Mayo clinic low-dose CT challenges of 2016 [141] and of 2021 [144] which encompass 30 and 300 patient scans respectively of roughly 70 slices each with noisy reconstruction and projection data simulated from clean reconstructed volumes. Secondly, the LoDoPaB-CT dataset [124] which uses 800 patient scans selected from the LIDC/IDRI database and contains over 40,000 scan slices. Thirdly, the IEEE ICASSP Grand Challenge 8 [20] which also utilizes the LIDC/IDRI database and contains 1010 3D cone-beam CT (CBCT) images. All three noise challenges rely on vendor reconstructed images that subsequently are backprojected to create corresponding projection data / sinograms which then are supplemented with simulated noise. Whereas the first two publications simulate their noisy data only by applying Poisson noise to the projection data, the third generates CBCT projection data with a custom noise simulator that accounts for photon counts, flat-fields, electronic sources, and detector cross-talk as sources of noise. Similar approaches have been undertaken by Bruno de Man et al. from GE research [45, 212] and Jingyan Xu and Benjamin M. W. Tsui [217] and shall be the basis for this work's noise simulator as well.



**Chosen noise simulation approach** In this work, we use a simplified version of the noise model used in XCIST [212]:

$$I_i = f_{CONV} \sum_k E_k \cdot \mathcal{P}(DQE_{ik} \cdot (A_{ik} + S_{ik})) + \mathcal{N}(\sigma_{electronic}) \tag{5.2}$$

$$I = \Gamma_{\sigma_{cross-talk}}[I_1, I_2, ..., I_I]^T \tag{5.3}$$

where $i$ is the pixel index of the detector $I$, $E_k$ is the energy level with energy index $k$, $A_{ik}$ are the incident photons in the pixel, $S_{ik}$ the scattered photons in the detector. $DQE_{ik}$ is the detector quantum efficiency and $f_{CONV}$ the energy to electron conversion rate. The noise process is described by $\mathcal{P}$, a Poisson random generator, and $\mathcal{N}(\sigma_{electronic})$ is a zero mean Gaussian random generator with standard deviation $\sigma_{electronic}$. Finally, $\Gamma_{\sigma_{cross-talk}}$ is a $\mathbb{R}^{D\times D}$ matrix that models detector cross-talk, defined as a fraction of the signal $\sigma_{cross-talk}$ that is shared between adjacent pixels.

This describes a full model of the detector behaviour given incident photons. In XCIST, the incident photons can be simulated by a Monte Carlo particle simulation based on a known source energy spectra and material decomposition of a sample. If the precise behavior of the energy-integrating detector is well understood for each energy level, the parameters $f_{CONV}$, $DQE_{ik}$, and $E_k$ can be incorporated. These parameters relate to the conversion of incident photons to measurements. However, for machine learning applications, the physics simulation would demand an unreasonably high computational time (several years for a sufficiently large dataset), necessitating simplifications of the model. In particular, the approximations done in this work assume that the measurement photons were produced by a monochromatic source ($k = 1$) and that there are no scattered photons measured ($S_k = 0$). Additionally, both the detector quantum efficiency of the pixels $DQE_k$ and the photon-to-electron conversion rate $f_{CONV}$ are assumed to be equal to one. As a detector specific calibration of these values is unknown and not easily obtainable without specialized lab equipment. This means that all photons reaching the detector are assumed to be measured and no loss of signal is present. These assumptions of course limit how close the simulation is to reality and following paragraph discusses their effects. The assumption of a monochromatic source is very common in the field of CT reconstruction and is the basis of the most commonly used version of the Radon transform. The effect of this assumption on the noise simulation is that there is no energy-dependent noise being added, but not that there is no noise. The assumption that no scattered photons are measured refers to omitting spatially dependent scatter which can be assumed to be low in comparison to the measured signal. However, the modeled Poisson noise itself still considers intensity-dependent scatter, as all noise in CT comes from photons that "attenuated", i.e. did not follow a straight path. Assumptions on the detector quantum efficiency of the pixels and the photon-to-electron conversion rate as well as the aforementioned two are simplifications that are made to limit the computational complexity of the noise simulation model. Adding noise that accounts for those effects requires is nowadays mainly available via Monte Carlo physics simulators like the mentioned XCIST software that we base our model on. This software has been used by GE Healthcare to validate their models for clinical implementation and can be assumed to be reasonably accurate. However, their computational footprint makes them infeasible to use for ML-sized datasets and requires us to limit the parameters of our noise simulation.

Thus, the chosen noise simulation approach to model the final measurement in the detector $I_D$ is:

$$I_i = \mathcal{P}(A_i) + \mathcal{N}(\sigma_{electronic}) \tag{5.4}$$

$$I = \Gamma_{\sigma_{cross-talk}}[I_1, I_2, ..., I_I]^T. \tag{5.5}$$

### 5.1.3 Training data

For the development of a ML-based denoising algorithm the most important element is adequate high-quality training data. In a supervised training framework that means

that there are pairs of input and target data. The algorithm is trained on these data pairs and learns a mapping from the input images to the target images. Zainulina et al. [223] concluded that such supervised deep learning methods show the best performance, but the requirement of paired images may not always be easy to accomplish. For the case of image denoising, this means noisy CT sinograms/reconstructions as an input and noise-free or "clean" CT sinograms/reconstructions as a target data.

Since the publication of the 2DeteCT dataset in 2023 [98], these paired images for supervised learned CT denoising are available. The 2DeteCT dataset comes with pairs of real measurements of the same object, one with near-zero noise, and one with high levels of noise. For this work, we deal with three types of sinograms: experimental clean sinograms that have been measured, experimental noisy sinograms that have been measured, and simulated noisy sinograms (based on the experimental clean data) that were simulated to contain the same level of noise as the experimental noisy sinograms. The corresponding acquisition parameters of this experimental data can be seen in Table 5.1).



In the remainder of this chapter, we will use the term "experimental noisy data" in

Table 5.1: Summary of the acquisition parameters of the 2DeteCT dataset, adapted from [98]. Mode 1 corresponds to the experimental low-dose, high-noise data; Mode 2 corresponds to the experimental high-dose, low-noise data. The Thoraeus filter is a compound filter made of Sn 0.1mm, Cu 0.2mm, Al 0.5mm. The SOD and SDD values are based on the motor readings of the FleX-ray scanner which get translated into physical quantities and are subject to alignment errors.

| **Acquisition parameter** | **Mode 1** | **Mode 2** |
|---|---|---|
| Tube voltage | 90.0 kV | 90.0 kV |
| Tube power | 3.0 W | 90.0 W |
| Filters used | Thoraeus | Thoraeus |
| Exposure time | 50.0 ms | |
| Binned detector pixel size | 149.6 µm | |
| Number of binned detector pixels | 956 | |
| Source to object distance (SOD) | 431.020 mm | |
| Source to detector distance (SDD) | 529.000 mm | |
| Number of projections | 3601 | |
| Angular increment | 0.1 deg | |

reference to raw low-dose CT measurement data acquired by a real-world experimental CT system. The term "simulated noisy data" will be used for artificially generated data for which artificial noise was applied to "clean" raw measurement data.

For experimental noisy data, the creation of corresponding image pairs requires a careful acquisition design to avoid that the algorithms would also learn a transformation or change of image content. The exact same CT slice needs to be scanned twice which makes it necessary to change the acquisition settings without infringing with the scanned object. The five main influencing acquisition parameters for the noise level within the CT images have been identified as source current ($I$), source voltage

($V$), exposure time ($t$), number of projections ($n_{proj}$), and number of averaged images ($n_{avim}$) [172]. Overall, the quantum noise in the reconstructed CT images is then inversely proportional to the square root of the number of detected photons. The aforementioned factors have their individual proportional influence on this number, which is given by: $\frac{V^{1.3}}{\sqrt{I \times t \times n_{proj} \times n_{avim}}}$

Analysing this formula, we can determine the relationship between the acquisition parameters and the corresponding noise level in the reconstructed CT slices. Since the used tube voltage $V$ not only influences the noise level but also changes the energy of the used X-ray photons, a change of this factor was omitted. The number of averaged images $n_{avim}$ could not be decreased further than one and since the scanner was already operated close to the shortest possible exposure time $t$, changing that parameter was also not feasible. To avoid artifacts due to insufficient sampling of the object we did not decrease the number of projections $n_{proj}$. Therefore, the tube current $I$ was the only feasible option to change and both the noisy and the clean CT scans were acquired with the exact same parameters except for the tube current. For the clean data this was 1000 µA whereas the noisy data had a 30 times smaller tube current of 33.3 µA.

For simulated noisy data creating corresponding image pairs is more straightforward. Given the "clean" data acquisition, a modification of the noise model in Equation 5.4 can be used to simulate artificial noise into the clean image. Given the noise-free incident photons $A_i$ and that the outcome of the Poisson process can be described as an addition $\mathcal{P}(A_i) = A_i + P_i$, where $P_i$ is just the noisy photons, we can rewrite Equation 5.4 for the acquisition of clean data as:

$$I_i^{clean} = A_i + P^{clean} + \mathcal{N}(\sigma_{electronic}), \tag{5.6}$$

where the assumption of $P^{clean} = 0$ can be made. This is not strictly true, but for a sufficiently large incident photon count $A_i$ it is approximately true. For the noisy acquisition thus the following holds:

$$I_i^{noisy} = A_i + P_i^{noisy} + \mathcal{N}(\sigma_{electronic}). \tag{5.7}$$

$$I_i^{noisy} = I_i^{clean} + P_i^{noisy} \tag{5.8}$$

$$I_D^{noisy} = I_D^{clean} + \Gamma_{\sigma_{cross-talk}}[P_1^{noisy}, P_2^{noisy}, ..., P_I^{noisy}]^T. \tag{5.9}$$

To appropriately simulate the low-dose measurements $I_D^{noisy}$, the noise distribution part of the total signal $P_i^{noisy}$ has to be produced, i.e. the Poisson component of the noise. Technically, $A_i$ would be a different number of photons for the clean and noisy images, as the noise mostly arises from the low photon count in our experiments and simulations. However, direct measurement of photon counts is not available and thus direct extraction of this noise from measured data is not possible. Therefore, the noise is parameterized by multiplying the flat-field corrected sinogram $ILI \in [0, 1]$ (see paragraph "Pre-processing") by a parameters corresponding to the number of photons in vacuum, $I_0$, and generating Poisson statistics from its result as

$$P_i^{noisy} = I_0 \cdot ILI_i^{clean} - \mathcal{P}(I_0 \cdot ILI_i^{clean}). \tag{5.10}$$

In this model, $I_0$ is the parameter to control the level of noise added to the clean data, a lower value representing noisier data. Based on the value in the XCIST software [212], a $\sigma_{cross-talk}$ of 5% of the signal is added.

### 5.1.4 Method development

The noise simulation and the algorithms for learned denoising in this work have been developed in LION (Learned Iterative Optimization Networks)[132], an open-source toolbox for learned tomographic reconstruction implemented in Python. With a designated data loader for the 2DeteCT dataset and with CT experiments set up in a reproducible way it serves as an environment for a standardized comparison of the methods described below.



For the learned denoising algorithms we selected two common convolutional neural networks (CNNs) for image processing tasks that have been used for both natural images but also for computed tomography images in particular: The generic U-Net [173] and the mixed-scale dense neural network (MSD-Net) [160]. The U-Net, originally developed for the segmentation of biomedical images, has been adopted in many fields as a baseline for image reconstruction based on neural networks. The MSD-Net has proven to be particularly effective for computed tomography [125, 159]. Its three main advantages are as follows: First, it has an advanced neural network architecture that uses dilated convolutions instead of traditional scaling operations to learn features at different scales. Second, it uses significantly fewer feature maps and trainable parameters which makes training it less computationally demanding and reduces the risk of over-fitting. Third, it has been applied to denoising large tomographic images and it has been proven that it can be easily applied to similar problems with minimal changes [160].

### 5.1.5 Comparison metrics

To evaluate the performance of the CNNs trained on either experimental or simulated noisy data, we consider two main comparison cases. In the first case, we test the performance of the algorithms in the setting that they have been trained on, i.e. settings in which they are supposed to work well. This means that if an algorithm is trained for denoising simulated noisy data this situation is used to score their overall performance. The same holds true for algorithms trained on experimental noisy data. For this we compare the output of our learned denoisers to the "clean" target data using the comparison metrics described below. In the second case, we want to compare the performance of the algorithms in settings for which they have not been trained for. This serves the purpose of checking their generalization to other tasks. It answers the question whether the algorithm generalizes to another noise model and its severity. In other words, whether the learned algorithms can also denoise input data without being trained on the specific noise of that data. This is particularly interesting for the case in which the learned denoisers are trained on simulated data and applied to experimental noisy data.

Using these comparison cases, we require comparison metrics with which we can evaluate the performance of the algorithms. Namely, how close the denoised images obtained from these algorithms are to the ground truth images. These metrics have to be able to measure two qualities: How well does the algorithm recover the structure of the imaged object from the noisy data? How well does it restore a good signal with respect to the overall noise in the reconstructed image?

Two commonly used metrics for these tasks are the structural similarity (SSIM) [202] and the peak signal-to-noise ratio (PSNR) [69]. The SSIM is a metric that indicates in a range from 0.0 to 1.0 how similar the compared image is to a ground truth, where 1.0 means they are identical. The PSNR is a metric that calculates the ratio between the highest attainable value of a signal and the strength of corrupting noise that impacts the fidelity of the image. Higher values in both metrics indicate a better algorithm performance. It is worth noting that these two commonly used quantitative metrics, may not be suitable for tomographic reconstruction or scalar fields [30, 31]. In reconstruction tasks such as CT imaging in medicine, PSNR and SSIM do not necessarily reflect a task-dependent better image [73, 136]. Therefore, it is suggested that evaluations consider such downstream tasks of the imaging rather than solely relying on traditional metrics. Additionally, the unbounded nature of CT images poses challenges for metrics like PSNR and SSIM, as the range of pixel values can vary. Different approaches to evaluating reconstruction performance, such as clipping or preserving the result range, can significantly impact reported performance. However, these metrics are still commonly used for a quantitative assessment of images. Since we are interested in measuring performance differences rather than rating the performance itself, they are also used in this work.

In our performance analysis we follow previous work by Zeng et al. [224] who argued that image artifacts due to beam hardening and photon-starvation are particularly difficult to evaluate meaningfully with quantitative metrics in the sinogram domain. They require a visual inspection in the reconstructed image domain. Therefore, we also include a qualitative, visualization-based evaluation between the results of denoised low-dose CT scans in the reconstructed image domain.

### 5.1.6 Computational experiments

For this work, we first applied the denoising in the projection domain, i.e. denoising beam intensity loss images ($ILI$), for three reasons: i) quality of denoising reconstructed images depends on the used reconstruction method; ii) artifacts caused by the noise in the projection domain are harder to remove after reconstruction; iii) noise in the projection domain is spatially uncorrelated. After evaluating the results of this approach we additionally trained denoising algorithms with an optimization in the reconstruction domain mapping directly from sinogram to reconstruction. For this, we included an FBP reconstruction in the pipeline of the models described below and visualized in Figure 5.1.

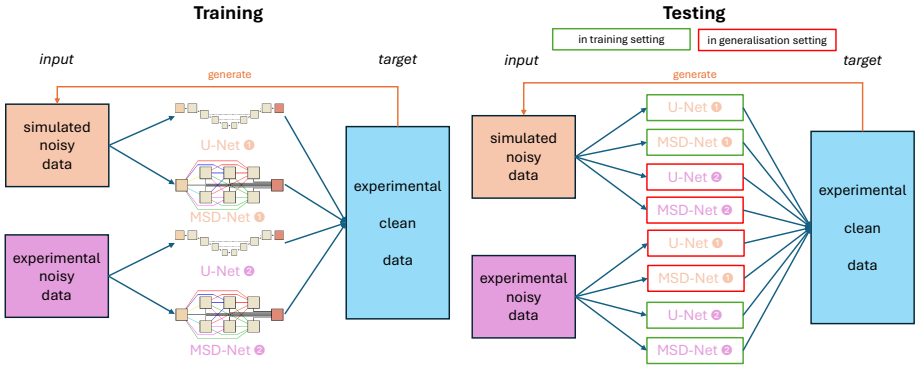



**Figure 5.1**: Training and testing scenarios for learned denoising networks (U-/MSD-Net illustrations adopted from [160]).

We prepared the training data for our learned denoisers by simulating noisy data from the "clean" experimental measurement data as described in Section "Training data" and using the unchanged clean data as ground truth target data. Consequently, there are two respective image pairs for supervised learning available: First, the simulated noisy data as an input and the experimental clean data as a target. Second, the experimental noisy data as an input and the experimental clean data as a target.

These image pairs were split into $\sim 80\%$ training data (3930 slices), $\sim 10\%$ validation (550 slices) and $\sim 10\%$ testing data (470 slices). Each algorithm was trained for 100 epochs using the Adam optimization algorithm [97]. The final model parameters were selected based on minimal validation loss. The computations were carried out on a GPU-server with 4x RTX 2080Ti (11GB), 384GB RAM, and 2x 16-core Xeon CPUs as well as a GPU-server with 2x RTX A6000 (48GB), 1TB RAM, and 2x 16-core Xeon CPUs.

After the training of the two neural network architectures on the two supervised learning settings, each of the four resulting trained networks was applied to their own test sets but also to the test sets of the data type they have not been trained on. A visual overview of this is given in Figure 5.1.

## 5.2 Results and discussion

### 5.2.1 Empirical selection of noise level

For our comprehensive study on the differences between the observed performances of algorithms trained on simulated noisy data and on experimental noisy data it was particularly important to have noise levels in our simulated noisy data that are representative of the noise levels present in our experimental noisy data. Therefore, we tried out various values of $I_0$ for our noise simulation approach and compared

both the resulting simulated noisy data and the experimental noisy data to the "clean" sinogram data with respect to PSNR and SSIM. Furthermore, the quantitative comparison was also carried out in the reconstruction domain, i.e. comparing the FBP-reconstructed images of the experimental and simulated noisy data to the "clean" reference reconstructions of the 2DeteCT dataset. Observing similar numerical values w.r.t. PSNR and SSIM for both experimental and simulated noisy data we can argue that our noise model generates simulated noisy data with a similar noise level. The results of this comparison can be found in Table 5.2. For all noise levels of the simulated noisy data the SSIM and PSNR values in the sinogram domain are significantly larger than the respective values for the experimental noisy data. A visual comparison of the different noise levels in the sinogram domain proofed uninformative as displayed in Figure 5.2.



Corresponding quantitative and qualitative analyses in the reconstruction domain showed similar image metrics for both the simulated and experimental noisy data. For a noise level of $I_0 = 200$ the PSNR value is closest to the same metric for the experimental noisy data, whereas the SSIM value shows its best agreement for a noise level of $I_0 = 300$. Since the task at hand is learned denoising, we chose to rely on the agreement with respect to the PSNR value and chose a noise level of $I_0 = 200$ for our computational experiments. A qualitative inspection of the images in Figure 5.3 agrees with this parameter choice.

A detailed inspection of Figure 5.3 furthermore showed a strong influence of the attenuation of the objects in each scan on the similarity between reconstructions based on simulated and experimental data. Simulated noisy image slices with no or only small objects with high attenuation (stones) appear to be visually close to the experimental noisy images. However, if those objects are bigger or grouped closely, the experimental noisy images show streaking artifacts caused by beam hardening, not visible in the reconstructions of the simulated noisy data. As previously mentioned, the noise model used in this work assumes mono-energetic photons and consequently cannot capture this behaviour. For the high-dose measurement data, which is the basis for the simulated noisy data, a high enough number of photons is detected and the reconstructed images do not present streaking artifacts due to beam hardening and some level of photon starvation.

### 5.2.2 Sinogram denoising

The quantitative analysis of the performance of the different CNNs trained on either experimental or simulated noisy data was carried out in both the sinogram and the reconstruction domain (FBP of model output) and is presented in Table 5.3. The evaluation in the sinogram domain shows that for both CNN architectures, U-Net and MSD-Net, the training on simulated noisy data performs better in both application cases, experimental and simulated noisy data. Applying the U-Net trained on experimental noisy data to simulated noisy data performs similarly well whereas the MSD-Net trained on experimental noisy data is not able to generalize well. Applying the U-Net trained on simulated noisy data to experimental noisy data yields a lower

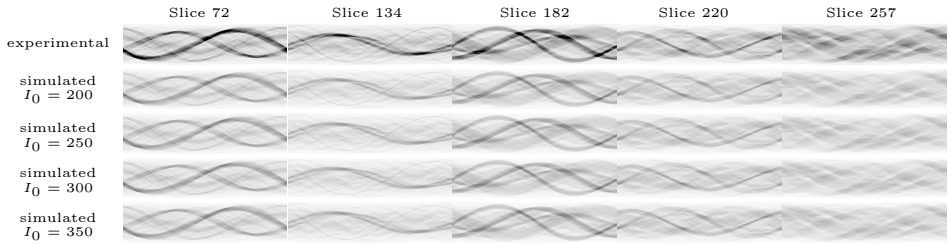



**Figure 5.2**: Visual comparison of the sinograms of experimental and simulated noisy data with different levels of $I_0$ (200, 250, 300, 350) from the 2DeteCT dataset for the slices with indices 72, 134, 182, 220, 257.



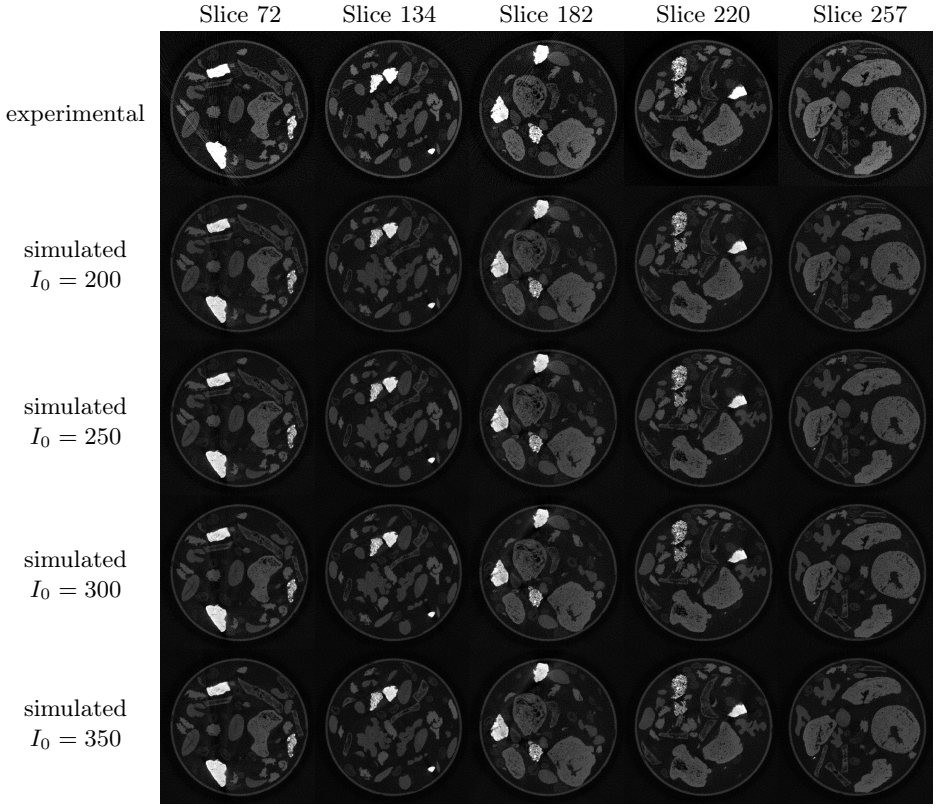



**Figure 5.3**: Visual comparison of the FBP-reconstructed images of the experimental and simulated noisy data with different levels of $I_0$ (200, 250, 300, 350) from the 2DeteCT dataset for the slices with indices 72, 134, 182, 220, 257.

Table 5.2: Empirical selection of the appropriate noise level $I_0$ to generate the simulated noisy training data based on the SSIM and PSNR values of the data with respect to the ground truth (wrt GT) data of "mode 2".

| **Type of Noisy Data** | **Noise Level** | **Evaluation in Sinogram Domain** | | **Evaluation in Reconstruction Domain (FBP of Noisy Data)** | |
|---|---|---|---|---|---|
| | | SSIM (wrt GT) | PSNR (wrt GT) | SSIM (wrt GT) | PSNR (wrt GT) |
| **Experimental Noise** | - | 0.2658 ± 0.0963 | 19.8130 ± 4.6583 | **0.1899 ± 0.0987** | **21.8024 ± 3.5996** |
| **Simulated Noise** | $\mathbf{I_0 = 200}$ | 0.2965 ± 0.0409 | 25.7190 ± 0.8567 | 0.1364 ± 0.0307 | **21.4468 ± 1.8307** |
| Simulated Noise | $I_0 = 250$ | 0.3448 ± 0.0435 | 26.6607 ± 0.8634 | 0.1627 ± 0.0356 | 22.3976 ± 1.8311 |
| **Simulated Noise** | $\mathbf{I_0 = 300}$ | 0.3854 ± 0.0451 | 27.4191 ± 0.8678 | **0.1865 ± 0.0397** | 23.1659 ± 1.8317 |
| Simulated Noise | $I_0 = 350$ | 0.4201 ± 0.0459 | 28.0517 ± 0.8725 | 0.2083 ± 0.0432 | 23.8089 ± 1.8325 |



performance of the learned denoiser on the test set images. The overall PSNR is 3dB lower and also the SSIM metric is 0.0522 lower than its application to simulated noisy test set data. For the MSD-Net this gap is even more significant. The MSD-Net trained on simulated noisy data applied to experimental noisy data yields a 21.7057 dB lower PSNR and a 0.1323 lower SSIM on the test set images compared to its application to simulated noisy test set data. This might be due to the much lower number of parameters of the MSD-Net which is not able to capture the experimental noise equally well as the simulated artificial noise.

However, CT reconstruction is an inverse problem that can exacerbate noise from the sinogram during the reconstruction process. Furthermore, applying the required sinogram pre-processing steps changes the nature of the noise model in a complex way. Therefore, evaluating the performance of the denoisers in the reconstruction domain is scientifically more relevant since even small errors in the sinogram domain might be larger in the reconstruction domain. For this reason, Table 5.3 also compares the performance of the sinogram denoisers in the reconstruction domain (FBP of model output).

In there we can observe that the high performance in denoising the sinograms does not carry over to the reconstruction domain. Both the structural similarity and the PSNR in this domain drop substantially. Additionally, the evaluation in the reconstruction domain shows that learned denoising of experimental noisy data performs best if the CNNs are trained on experimental noisy data, as it is expected. Furthermore, the U-Net architecture seems to pick up the image content in terms of structural similarity (SSIM) better than the MSD-Net when trained on experimental noisy data. The PSNR performance is better for the MSD-Net in all training settings except for the case of training on simulated noisy data and testing on experimental noisy data.

After the uninformative visual inspection of the simulated noise in the sinogram domain, and considering that the ultimate goal is to obtain better reconstructed images, the qualitative analysis of the model performances was only carried out in the reconstruction domain which can be found in Figure 5.4. The qualitative visual inspection, also in comparison to the reference images displayed in Figure 5.5, shows that the models for sinogram denoising (found within the first four rows of the

Table 5.3: Quantitative performance analysis with PSNR and SSIM of the differently trained models in the reconstruction domain for the two different testing data with respect to the ground truth data from the iterative reference reconstructions of "mode 2" from the 2DeteCT dataset.

| **Method** | **Training Data** | **Metric** | **Testing Data** | |
|---|---|---|---|---|
| | | | Experimental Noisy Data | Simulated Noisy Data |
| **Evaluation in Sinogram Domain** | | | | |
| U-Net (2) | Experimental Noisy Data | SSIM | 0.8126 ± 0.0194 | 0.8167 ± 0.0199 |
| | | PSNR | 18.4966 ± 0.6278 | 19.3181 ± 0.5575 |
| U-Net (1) | Simulated Noisy Data | SSIM | 0.8273 ± 0.0240 | 0.8795 ± 0.0206 |
| | | PSNR | 33.4602 ± 0.9533 | 36.6016 ± 0.5616 |
| MSD-Net (2) | Experimental Noisy Data | SSIM | **0.8613 ± 0.0211** | 0.8239 ± 0.0216 |
| | | PSNR | **36.2182 ± 0.7214** | 20.4747 ± 0.6793 |
| MSD-Net (1) | Simulated Noisy Data | SSIM | 0.7512 ± 0.0226 | **0.8835 ± 0.0198** |
| | | PSNR | 16.3208 ± 1.3965 | **38.0265 ± 0.7412** |
| **Evaluation in Reconstruction Domain (FBP of model output)** | | | | |
| U-Net (2) | Experimental Noisy Data | SSIM | **0.6134 ± 0.0732** | 0.6273 ± 0.0717 |
| | | PSNR | 26.7127 ± 1.9780 | 27.5290 ± 1.9405 |
| U-Net (1) | Simulated Noisy Data | SSIM | 0.5504 ± 0.0677 | 0.6351 ± 0.0713 |
| | | PSNR | 28.3307 ± 2.0810 | 32.5568 ± 2.0169 |
| MSD-Net (2) | Experimental Noisy Data | SSIM | 0.5984 ± 0.0741 | 0.6152 ± 0.0723 |
| | | PSNR | **30.9185 ± 1.9707** | 28.3031 ± 1.9314 |
| MSD-Net (1) | Simulated Noisy Data | SSIM | 0.3854 ± 0.0469 | **0.6372 ± 0.0469** |
| | | PSNR | 11.6366 ± 2.3636 | **32.6552 ± 2.0173** |
| **Evaluation in Reconstruction Domain (model output)** | | | | |
| FBP+U-Net (2) | Experimental Noisy Data | SSIM | **0.8161 ± 0.0592** | 0.7466 ± 0.0681 |
| | | PSNR | 29.8398 ± 2.1834 | 28.0435 ± 2.0848 |
| FBP+U-Net (1) | Simulated Noisy Data | SSIM | 0.5957 ± 0.0844 | 0.6693 ± 0.0820 |
| | | PSNR | 26.8841 ± 3.4800 | 28.8134 ± 3.9418 |
| FBP+MSD-Net (2) | Experimental Noisy Data | SSIM | 0.7829 ± 0.0749 | 0.7892 ± 0.0731 |
| | | PSNR | **32.0684 ± 1.9309** | 32.0704 ± 1.9580 |
| FBP+MSD-Net (1) | Simulated Noisy Data | SSIM | 0.7615 ± 0.0702 | **0.8204 ± 0.0567** |
| | | PSNR | 30.6211 ± 2.0626 | **33.1053 ± 2.0120** |

figure) do not produce high-quality reconstructions, particularly regarding fine image features/details. The images exhibit lower noise than the FBP reconstructions of the noisy data directly, but there is a noticeable loss of image sharpness.

### 5.2.3 Optimization in the reconstruction domain: mapping directly from Sinogram to Reconstruction

Having observed that a good model performance in the sinogram domain does not necessarily carry over to the reconstruction domain we wondered whether training the denoising algorithms with an optimization in the reconstruction domain mapping directly from noisy sinograms to "clean" reconstructions, would prove more effective as well. The model performance w.r.t. clean target reconstructions can be found in the bottom third of Table 5.3 and in the bottom half of Figure 5.4. The relative performance of the networks for the respective combinations of training and testing settings is the same as before, but the results are substantially better. We observe an increase of 0.2027 in the SSIM for the best performing model in the constellation experimental noisy training data and experimental noisy testing data and an increase of 0.1832 in the SSIM for the best performing model in the constellation simulated noisy training data and simulated noisy testing data. Also the performance with respect to the PSNR for each corresponding constellation of training and testing data is better if the models are optimized in the reconstruction domain.

A qualitative analysis of the images in Figure 5.4, also in comparison to the reference images displayed in Figure 5.5, show that the performance drop of training on simulated noisy data but testing on experimental noisy data is more substantial than what the performance metrics would suggest, as these metrics capture global performance rather than local. In all of the slices inspected, this particular train/test case produces the worst images of the quadruplet, for both models. Increased "graininess" permeates the entire image, and the low-intensity objects appear more porous than expected.

## 5.3 Discussion and conclusions

In this work, we aimed to answer the question to which extent algorithms trained on simulated noisy data are applicable to real-world experimental noisy data. This was achieved through the implementation of a realistic yet computationally efficient simulation method and utilizing less-commonly available raw experimental measurement data.

After tuning the noise simulation to the experimentally measured noise level, our empirical selection of $I_0$ to set the noise level proved to be an adequate choice both in the qualitative and quantitative assessment (PSNR and SSIM) in the reconstruction domain.

Differences in the simulation were mainly observed in the presence of large or closely grouped high-attenuation samples in the respective image slices, i.e. when beam

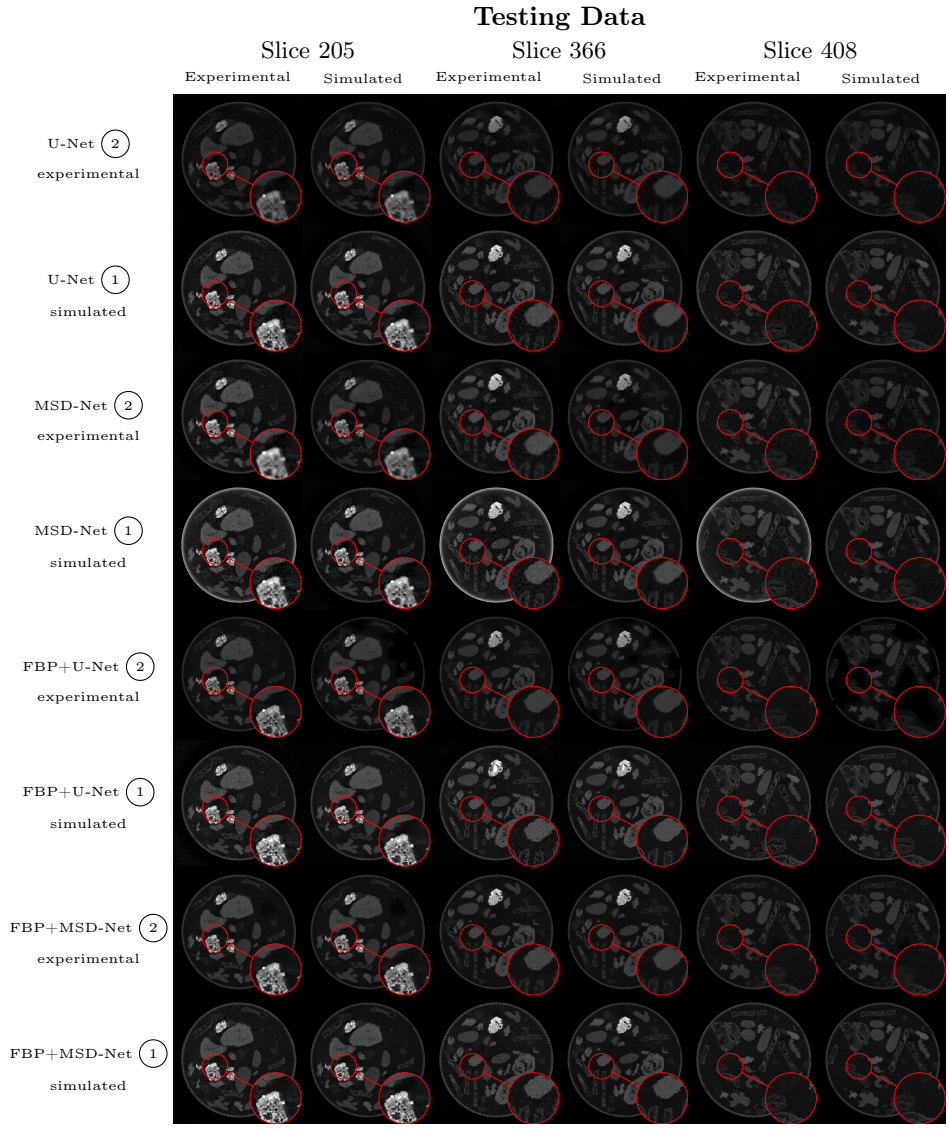



**Figure 5.4**: Qualitative performance analysis of the differently trained models in the reconstruction domain for the two different testing data (slices indices 205, 366, 408).

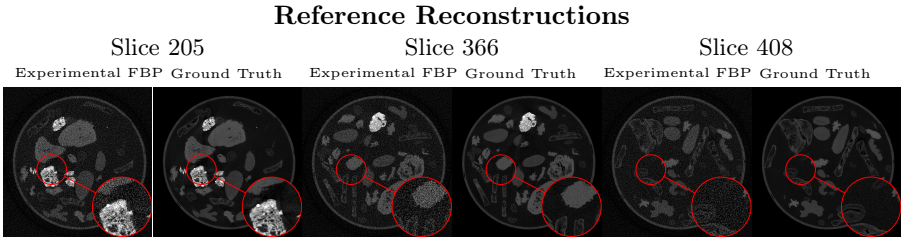



**Figure 5.5**: Reference reconstructions for the qualitative performance analysis of the differently trained models in the reconstruction domain for the two different testing data (slices indices 205, 366, 408).



hardening is present. This is expected, as the chosen noise model for simulation assumes monochromatic sources and thus cannot simulate highly non-linear effects such as beam hardening.

While sinogram denoising achieved better results with simulated noisy data when evaluated in the sinogram domain, the performance did not carry over to the reconstruction domain where training on experimental noisy data showed a higher performance in denoising experimental noisy data. As previously mentioned, this is caused by the inherent ill-posedness of CT reconstruction, that amplifies any remaining noise in the process. Therefore, training the denoising algorithms with an optimization in the reconstruction domain mapping directly from sinogram to reconstruction showed significant improvements in model performance. This is especially noticeable in terms of structural similarity and qualitative visual inspections of the reconstructions. It seems that the artifacts introduced by the FBP reconstruction are not too severe to mitigate via the subsequent post-processing network.

Our findings highlight the importance of carefully designing a noise simulation approach and choosing appropriate noise levels that match experimental data well. If possible the training should be conducted with an optimization in the reconstruction domain, i.e. mapping from raw measurement data to desired target reconstructions. In machine learning for computational imaging, simulated data can be quite different from experimental data, which can impact the transfer of learned systems to the real-world. In particular, the distributions of the training and testing data should be as close as possible and therefore training on experimental noisy data, if available, is preferable when the models are subsequently applied to experimental data. In our experiments, models trained on simulated data exhibit a measurable quantitative performance drop from simulated noisy testing data to experimental noisy testing data. This is even more noticeable by qualitative visual inspection, because these models produce the noisiest images from all the cases.

Ultimately, this research shows that appropriately simulating real noise is important in learned CT research. While computationally fast noise models, like the one presented in this work, will produce data that are close enough to experimental data to make

the models transferable to real-world applications, a drop in performance is expected. Hence, it is advisable to utilize real-world experimental data for training learned denoisers whenever feasible. Furthermore, one should be cautious with, presenting performance outcomes solely based on simple performance metrics when training only on simulated noisy data. As discussed before, our simulation model already captures much of the complexity of the experimental noise in the measurements. However, this work shows that the non-linearity of the imaging process is not captured well enough and that future work should investigate computationally efficient ways of including effects such as beam hardening or photon starvation. Possibly, generative models trained on experimental noisy and "clean" data could solve this challenge or alternatively simplified Monte Carlo particle simulations could be investigated. This study can serve as a starting point for crafting and testing even more sophisticated noise simulation approaches that might be able to close the sim-to-real gap [152, 205] for CT image denoising.

## 5.4 Code availability

Python scripts for setting up the neural network training as well as the evaluation of the noise reduction performance in the way described above are published on GitHub: https://github.com/CambridgeCIA/LIONscripts/paper_scripts/noise_paper. They make use of the ASTRA toolbox [1, 154, 190], which is openly available on ( www.astra-toolbox.com ) and tomosipo [81].