Title page

Dose Reduction and Image Enhancement in Micro-CT using Deep Learning

Comment [AU1]: Please note that Supplementary Figures S1--S10 and Supplementary Tables S1--S3 are not cit in text.

Running Title: Deep Learning for Low Dose Micro-CT

Authors:

Florence M. Muller[1*]

Jens Maebe[1]

Christian Vanhove[1]

Stefaan Vandenberghe[1]

[1] Medical Image and Signal Processing (MEDISIP), Department of Electronics and Information Systems, Faculty of Engineering and Architecture, Ghent University, 9000 Ghent, Belgium

*Correspondence:

FlorenceMarie.Muller@UGent.be

+32 (9) 3320054

Mailing address: INFINITY Lab, Corneel Heymanslaan 10, Entrance 21, UZ Ghent Campus, 9000 Ghent, Belgium

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/mp.16385.

Abstract

Background. In preclinical settings, micro-computed tomography (CT) provides a powerful tool to acquire high resolution anatomical images of rodents and offers the advantage to *in vivo* non-invasively assess disease progression and therapy efficacy. Much higher resolutions are needed to achieve scale-equivalent discriminatory capabilities in rodents as those in humans. High resolution imaging however comes at the expense of increased scan times and higher doses. Specifically, with preclinical longitudinal imaging, there are concerns that dose accumulation may affect experimental outcomes of animal models.

Purpose. Dose reduction efforts under the ALARA (as low as reasonably achievable) principles are thus a key point of attention. However, low dose CT acquisitions inherently induce higher noise levels which deteriorate image quality and negatively impact diagnostic performance. Many denoising techniques already exist, and deep learning (DL) has become increasingly popular for image denoising, but research has mostly focused on clinical CT with limited studies conducted on preclinical CT imaging. We investigate the potential of convolutional neural networks (CNN) for restoring high quality micro-CT images from low dose (noisy) images. The novelty of the CNN denoising frameworks presented in this work consists of the model training using image pairs with realistic CT noise present in both the input and target image; a noisier image acquired with a low dose protocol at the input is matched to a less noisy image acquired with a higher dose scan of the same mouse.

Methods. Low and high dose *ex vivo* micro-CT scans of 38 mice were acquired. Two CNN models, based on a 2D and 3D four-layer U-Net, were trained with mean absolute error (30 training, 4 validation and 4 test sets). To assess denoising performance, *ex vivo* mice and phantom data were used. Both CNN approaches were compared to existing methods, like spatial filtering (Gaussian, Median, Wiener) and iterative total variation image reconstruction algorithm. Image quality metrics were derived from the phantom images. A first observer study (n=23) was set-up to rank overall quality of differently denoised images. A second observer study (n=18) estimated the dose reduction factor of the investigated 2D CNN method.

Results. Visual and quantitative results show that both CNN algorithms exhibit superior performance in terms of noise suppression, structural preservation and contrast enhancement over comparator methods. The quality scoring by 23 medical imaging experts also indicates that the investigated 2D CNN approach is consistently evaluated as the best performing denoising method. Results from the second observer study and quantitative measurements suggest that CNN-based denoising could offer a 2-4x dose reduction, with an estimated dose reduction factor of about 3.2 for the considered 2D network.

Conclusions. Our results demonstrate the potential of DL in micro-CT for higher quality imaging at low dose acquisition settings. In the context of preclinical research, this offers promising future prospects for managing the cumulative severity effects of radiation in longitudinal studies.

Keywords: Convolutional neural networks, Deep learning, Dose reduction, Image denoising, Micro-computed tomography

1. Introduction

Computed tomography (CT) uses X-rays to image the internal structure of objects by taking cross-sectional (2D) transmission profiles at different angles that are then reconstructed into a three-dimensional (3D) volume. In preclinical settings, micro-CT provides a powerful tool to acquire high resolution anatomical images of rodents and to *in vivo* non-invasively assess disease progression and treatment efficacy. This reduces and refines the usage of animals whereby each subject can serve as its own control during longitudinal follow-up studies [1,2].

Much higher resolutions are needed to achieve scale-equivalent discriminatory capabilities in rodents as those in humans [3]. High resolution imaging however comes at the expense of increased scan times and higher doses, which is often not desirable given the effects of radiation exposure [4]. Especially in longitudinal studies, when the same animal population is imaged at multiple time points, accumulated dose may affect experimental outcomes of animals models [1]. While typical radiation levels from micro-CT acquisitions (10-500 mGy per scan) are normally non-lethal to the animal (below the 6 Gy threshold of acute tissue damage), they can be substantial enough to impact biological pathways [5]. Note that like in clinical imaging, micro-CT is often used in combination with micro-PET (Positron Emission Tomography) and micro-SPECT (Single-Photon Emission Computed Tomography) to provide CT-based attenuation correction and anatomical reference maps. PET and SPECT scans however also deliver dose which can be higher than CT radiation [6]. Dose reduction efforts under the ALARA (as low as reasonably achievable) principles are thus important . However, lowering the radiation dose by decreasing tube current and/or shortening exposure time inevitably leads to photon starvation which reduces the signal-to-noise ratio (SNR) in the images. Noise negatively impacts image quality and diagnostic performance.

Various techniques have been developed to reduce noise of low dose CT scans, and these methods are either applied in the (i) projection-domain [7,8], (ii) image reconstruction process [9,10], or (iii) image-space [11,12]. Deep learning (DL) has become increasingly popular for medical imaging enhancement tasks, such as denoising, super-resolution and artefact removal [13,14]. In particular, convolutional neural networks (CNN) show important performance gains in terms of noise reduction in low dose imaging [15,16]. Given their ability to learn high-level features from pixel data through hierarchical networks, supervised CNN algorithms attempt to find a mapping function that reduces noise in low dose images (or also for limited angle tomography scans) from matching high dose images (or full angle acquisitions). Unsupervised DL methods for CT image restoration have also been explored [17,18] and

most derive from the deep image prior proposed by Ulyanov *et al.* [19] who showed that an image could be enhanced without requiring any prior training data other than the image itself.

Chen *et al.* [20,21] were one of the firsts to introduce a CNN for noise reduction in low dose CT. Later, Kang *et al.* [22] developed a deep CNN that combined a residual encoder-decoder network with wavelets to enable better retention of textural details. A 3D version of the basic ResNet structure was introduced by Yang *et al.* [23] with the aim to maintain spatial association between tissues and organs. Wolterink *et al.* [24] were the first ones to propose a generative adversarial network (GAN) for denoising of low dose CT images. Based on the adversarial learning method and the cost function (usually multi-objective functions) different variants exist with Wasserstein GAN and cycle-GAN being the two most studied models for image denoising of low dose CT [25-28]. For clinical low dose CT, a multitude of DL denoising models have been proposed, and each investigates different combinations of network architectures and loss functions [14-16]. Note that, besides image-to-image DL frameworks that serve as post-reconstruction tool, alternatives are to train the DL denoising model in the projection space or to develop an end-to-end DL algorithm that directly maps a noisy sinogram to a clean image (DL-based image reconstruction) [29,30].

While the principles of DL denoising are equally applicable to clinical and preclinical settings, most research efforts are spent on clinical low dose CT and only a few studies investigate the merits of DL-based image enhancement for preclinical imaging. Chen *et al.* [31] proposed a conditional GAN as framework for denoising micro-CT in the projection domain, and low dose micro-CT scans were artificially created by adding Poisson noise into the high dose projections. Yao *et al.* [32] developed an eight-layer asymmetric perceptual convolutional network for micro-CT image denoising. Clark *et al.* [33] trained a CNN to denoise 3D cardiac micro-CT data in the image-domain. For the latter two studies, the low dose image was reconstructed from under-sampled projection data of the standard dose micro-CT acquisition.

Preclinical data is a crucial component in medical research to answer biological/clinical questions - translational science brings preclinical knowledge (bench) to clinical practice (bedside), and vice versa. Our study aims to explore the potential and feasibility of DL to predict high quality micro-CT images from noisy images acquired at lower dose settings. We investigated two different CNN approaches and both CNN-based denoising models were trained on a dataset consisting of image pairs with realistic CT noise present in the input and target image that were obtained from low (LD) and high dose (HD) micro-CT acquisitions, respectively. While our work is certainly not the first to study the potential of neural networks for low dose image denoising, it stands by the use of real experimental data to define both the input and target image used for model training. In comparison, most training schemes described in clinical CT literature are based on the acquisition of a high dose

scan and the artificial creation (simulation) of a low dose image by adding noise or under-sampling the projection data. Of course, the acquisition of subsequent LD and HD scans of the same subject (here: rodents) is more easily feasible in preclinical set-ups.

2. Method

2.1. Data Acquisitions

All micro-CT data was obtained from the X-CUBE (Molecubes, Ghent, Belgium), a benchtop micro-CT system [34]. In total 38 mice were scanned by *ex vivo* whole-body spiral micro-CT. Acquired projection datasets were reconstructed iteratively (GPU-based ISRA) with 200 µm voxel size. Note that 200 µm voxels were chosen because of practical storage size and reconstruction time, but also, although the spatial resolution of the X-CUBE system is in the 50 µm range, it is quite common to reconstruct micro-CT images at 200 µm resolution especially for multi-modality imaging with preclinical CT. Two different scan settings, General-Purpose (GP) and High-Resolution (HR) protocols predefined on the X-CUBE, were selected to acquire the LD and HD scans, respectively. Table 1 compares the scan parameters for both protocols: The HR protocol delivers 10x more dose to the animal and deposits 3.5x more dose for image quality equivalence to the GP protocol. To guarantee spatial alignment between LD and HD images, MRtrix3 (https://www.mrtrix.org/) was used to perform rigid non-linear coregistration by applying the *population_template* command.

Besides the *ex vivo* mice scans, two Quality Control (QC) phantoms were scanned (with the same acquisition settings) for further evaluations. The Molecubes CT QC phantom is a cylindrical-shaped phantom that consists of: (i) nine contrast spheres with different diameter sizes, and (ii) a uniformly-filled part. The PlastiMouse (SmART Scientific Solutions) is a plasticized *ex vivo* version of a real mouse that can be used as a phantom. For each phantom, ten subsequent HD scans were acquired (10x HR) and averaged to obtain a ground truth ("noise-free") image. Similar experiments were repeated with the GP protocol to acquire images at five representative dose levels: 1x, 2x, 4x, 6x and 8x LD (e.g. 4x LD image corresponds to an acquisition at twice the radiation dose of 2x LD).

2.2. Network Architecture and Training Procedure

Important to recall is that the CNN denoising models were trained using image pairs with realistic CT noise present in both the input and target image; a noisier image acquired with a LD protocol at the input was matched to a less noisy image obtained with a higher dose scan of the same mouse. This

study investigated two CNN approaches, a 2D U-Net (referred to as 2D-Net1) and a 3D U-Net (abbreviated as 3D-Net2), where each method was implemented with a different training protocol as further detailed below. Both architectures (Figure 1) adopt an end-to-end fully convolutional network based on a four-layer U-Net structure [35]. Skip connections copy and concatenate the contraction layer output in the channel dimension with the expansion layer input. Both CNN approaches were trained on 30 (=80%), validated on 4 and tested on 4 *ex vivo* micro-CT scans. Random vertical and horizontal flipping was used for data augmentation. A pixel wise (voxel wise for 3D) loss function using mean absolute error was compiled with Adam optimizer [36] with a learning rate of 1E-4 and a batch size of 16. The 2D-Net1 was trained on 12,000 slices (400 slices per mouse with 30 mice in the training set). Image slices were not divided into smaller patches, but instead the model loaded the complete 2D slice into the network. Also note that the 2D-Net1 was presented with three adjacent slices (3-channel at the input layer) from which it created 64 feature maps thereby allowing to capture some coordinating spatial dependencies across adjacent slices (with the aim to minimize structural information loss). The training inputs were reduced in matrix size so to only contain relevant signal (remove unnecessary background). The 2D-Net1 was trained for 25 epochs (early stopping) and total training time was approximately seven hours. In comparison, the 3D-Net2 was trained on a patch-by-patch approach with one channel at the input layer and it used 100 random patches of 32x32x32 voxels in size per mouse for each training epoch (which equates to 3,000 patch units in total). Note that the size of the patch units was constrained by memory limits. Patch locations were arbitrarily defined around voxels of value above -300 HU to ensure that it comprised sufficient image content. For testing (during inference), the entire 3D image was passed through the network. The 3D-Net2 was trained for 22 epochs (early stopping) and total training time was about 18 hours. We also report the inference time of the 2D-Net1 and 3D-Net2 methods to process a mouse scan (for an image size of 500x370x370 voxels) which takes 41.2 and 56.8 seconds, respectively, when run on a NVIDIA Volta V100 graphics processing unit (32GB GPU memory).

2.3. Performance Evaluations

2.3.1. Comparator Methods

To compare the CNN denoised images to other (more standard) denoising techniques, three spatial filters and an iterative image reconstruction method using total variation (TV) minimization [38] were applied on the LD scans. Gaussian filter used a smoothing kernel with standard deviation of 0.5 voxels. Median filter applied a 3x3x3 kernel. Wiener filter used a smoothing kernel in the mean and variance of 3x3x3 neighboring voxels. The reconstruction system on the X-CUBE provides the option to include a statistical noise reduction algorithm that implements TV regularization to ISRA. The

acquired LD projection datasets were thus also reconstructed with ISRA-TV (with hyperparameter β of 0.0016 using 10 iterations in combination with 18 subsets) at 200 µm voxel size.

2.3.2. Quantitative Measurements of Image Quality Metrics on Phantoms

Noise was characterized by measuring the average standard deviation in CT number in five regions-of-interest (ROI) in the uniformity section of the Molecubes QC CT phantom. The contrast-to-noise (CNR) was quantified for the large (10 mm diameter) and smaller (4 mm diameter) sphere (see Section 1 in Supplementary_Material.doc). These quantitative analyses were conducted in AMIDE software. Root mean squared error (RMSE), peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) were quantified on the PlastiMouse phantom where the ground truth was available (acquired by averaging 10x HD). The quantitative metrics were measured between the denoised and ground truth on the entire 3D dataset of the PlastiMouse.

Additionally, to include a contrast-dependent measurement, linearity of the 2D-Net1 and 3D-Net2 denoising solution were determined using a linearity plate with five vials. One vial was filled with water (= 0 $mg.ml^{-1}$) and four vials were prepared with iodine solutions of different concentrations of contrast agent: 10, 20, 30 and 40 $mg.ml^{-1}$. The relationship between signal intensity (mean CT measured in a cylindrical ROI centrally placed on each vial) and respective iodine concentration was determined by linear regression analysis. Lastly, to complete the quantitative part on image quality assessment, we included an image resolution evaluation with a modulation transfer function (MTF) that was determined using a custom-made phantom. This phantom consisted of five glass capillary tubes filled with air (see Section 2 in Supplementary_Material.doc) and was scanned with GP and HR protocol. To assess the MTF, a line profile was drawn across the transverse section of the capillary tube phantom, showing bright and dark intensity modulations. In this way, the impact of the 2D-Net1 and 3D-Net2 denoising model on image resolution could be evaluated.

2.3.3. Multi-Observer Studies for Image Quality Assessment and Dose Reduction Estimation

Observer studies were included as an evaluation method to qualitatively compare the performance of the CNN algorithms against that of other denoising methods. The observer cohort consisted of international preclinical researchers and nuclear medicine physicians, working with (micro-)CT data on a daily-weekly basis. A first study (n=23) was set-up to score the overall quality of images processed by the different methods using an eight-point Likert scale (with 8 the highest score). In a second observer study (n=18), the 2D-Net1 denoised images were compared to images acquired at

different dose levels on the PlastiMouse. The subjective image quality assessment by the observers was used to estimate by how much the 2D-Net1 model could reduce dose based on how well it improved quality as ranked relative to the images from the different dose levels (obtained from LD scans). To ascertain the generalizability of both observational studies, the inter-observer agreement was assessed using Fleiss' Kappa [39], an extension of Cohen's Kappa for cases with more than two observers.

3. Results

3.1. Visual Comparison of Denoising Results

Representative (denoised) image slices of the *ex vivo* micro-CT test datasets are presented in Figure 2, 3 and 4 for visual comparative inspections. More examples as well as denoised phantom images are included in Section 3 of the Supplementary_Material.doc. All applied methods suppressed image noise to various extents. The three spatial filters left some grainy and streak image noise without elimination and blurred the structures at the ribcage bones and vertebrae. ISRA-TV and both CNN approaches removed most image noise and flattened the image appearance. Despite the smoothening looks, the 2D-Net1 and 3D-Net2 models showed superior performance in terms of noise suppression, but also in regards to retaining structural details (no blur introduced). This is supported by the subtraction images (Figure 3) obtained from the denoised images relative to the LD. In comparison, the subtraction images of both 2D-Net1 and 3D-Net2 methods (Figure 3(e) and (f)) present mostly the removed noise content and minimal differences on structural content.

The zoomed-in images in Figure 4 illustrate that the LD image was severely corrupted by noise and suffered from quality deterioration, which limits the detectability of tissue contrast in the abdomen. Both CNN approaches clearly demonstrated the best low contrast recovery from the noisy LD image. In comparison, denoising by spatial filtering retained a good contrast resolution but the overall visibility of smaller contrast regions was hampered due to limited noise removal and inherent resolution loss. ISRA-TV blurred the lower contrast regions and the introduced blur significantly compromised the resolution (sharpness) of structural details along the spine.

3.2. Quantitative Evaluations of Image Quality

The first column of Table 2 reports the noise levels measured on the uniform part of the Molecubes CT QC phantom. All methods resulted in considerable noise reduction relative to the LD image as

inferred from the lower mean value and smaller standard deviation in CT number. ISRA-TV, 2D-Net1 and 3D-Net2 gave the largest extent of denoising. As shown in Figure 5, Gaussian, Median and Wiener filtering led to an increase in CNR relative to the LD image, but appreciably larger improvements in CNR on both sphere sizes were found with ISRA-TV and CNNs. RMSE, PSNR and SSIM between the denoised and the ground truth image were derived for the PlastiMouse (Table 2). With respect to the LD image, all three metrics were notably improved by both CNN algorithms (i.e. lower RMSE, higher PSNR, higher SSIM) with slightly better results measured for the 2D-Net1. Figure S-13(b) presents the linearity results that provide a contrast-dependent evaluation of the 2D-Net1 and 3D-Net2 methods for different iodine solutions. Linear regression analysis showed a significant correlation between signal intensity and increasing contrast (i.e. iodine concentrations) which was coherent in case of the LD, 2D-Net1 denoised, 3D-Net2 denoised and HD image. As for quantitative image resolution assessment, Figure S-14 (see Supplementary_Material.doc) presents the results from the MTF calculations conducted on the custom-made capillary tube phantom. The MTF values derived from the LD, 2D-Net1 denoised, 3D-Net2 denoised and HD image showed a relatively good agreement, suggesting that CNN denoising solutions do not affect spatial resolution.

3.3. Quantitative Estimation of Dose Reduction

RMSE, PSNR and SSIM were also measured between the ground truth and images from different dose levels (obtained by averaging LD scans: 2x, 4x, 6x and 8x LD). Results are shown in Figure 6 by the black curve with black dots, which in fact exemplifies image quality improvement (i.e. decrease in RMSE, increase in PSNR and increase in SSIM) as a function of the number of averaged LD scans (indicative of higher doses). The metrics from Table 2 were matched on the black curve in Figure 6 (as shown by the colored dotted lines) to estimate the number of LD scans needed to get to the same image quality, which gives an estimate by how much dose could potentially be reduced (Table 3). The 2D-Net1 can achieve a 3.2 ± 0.5 times dose reduction, while the 3D-Net2 can offer a dose reduction factor of about 2.1 ± 0.2. To validate our methodology, the estimated dose reduction of the HD (3.9 ± 0.4) was compared to the known imaging dose difference of 3.5 between the GP and HR protocols used to acquire the LD and HD image, respectively.

3.4. Observer Studies

Results of the first observer study (n=23) (see Figure 7(a)) showed that the LD image had the lowest score and all applied denoising methods improved the scores. The 2D-Net1 and 3D-Net2 provided

substantially higher scores; the 2D-Net1 ranked on top with the best image quality score and associated smallest standard deviation (7.40 ± 0.35). Figure 7(b) presents the results of the second observer study (n=18). Using linear interpolation between the scores of 2x and 4x LD, the 2D-Net1 quality score gives a dose reduction factor of 3.00 (±0.23). Note that this result agrees well with the dose reduction factor (3.2 ± 0.5) estimated from the quantitative analyses. Statistical results for the inter-observer agreement assessment are presented in Section 5 of the Supplementary_Material.doc.

4. Discussion

One of the primary image degrading factors in medical imaging is noise. Physical process randomness, detector performance and limitations imposed by scanner design inevitably lead to the creation of quantum noise which reflects a level of variability in pixel (voxel) data over the image space. The problem with noise present in CT images is that it obscures relevant image content and limits the discrimination between soft tissue regions because of degraded image resolution and contrast. Restoring a noisy image to a higher quality is challenging: to distinguish actual noise content from true (image) signal is difficult, since edge and texture have, similar to noise, also high frequency components. When a spatial low-pass filter is applied to attenuate high frequency noise, it leads to image resolution loss (as also shown in our results). This is where DL denoising has shown promising potential to overcome the hurdles faced with low dose CT data.

While the principles of using DL for noise reduction (image enhancement) are equally applicable to clinical and preclinical settings, the large majority focuses on clinical low dose CT and very little work is conducted on the merits of DL for image denoising of low dose micro-CT. To the best of our knowledge, this study is one of a few [31-33] to investigate the feasibility and potential of DL denoising in micro-CT (preclinical) imaging and we here highlight the elements that differentiate this study from others. First, the DL noise reduction model was formulated on the basis of pairs of noisy images (that share the same underlying, noise-free image) instead of using a ground truth as target or a simulated low dose image as input. That is to say, both networks (2D-Net1 and 3D-Net2) learned the mapping from low to high dose images based on a training strategy relying on real-world data; the LD and HD images were acquired with two different scan protocols, so that high and low levels of realistic CT noise were present in the image pairs. In general, low dose acquisitions are "artificially" obtained by either adding photon noise to the projections or reducing the number of projections, but this does not necessarily represent the real-world situation [40]. Our methodology stands by the use of real experimental data for training as well as evaluating both CNN methods. This approach is

indisputably more suited (feasible) in preclinical settings compared to clinical imaging situations with humans. For the evaluations, multiple subsequent scans were done at different dose levels to derive dose reduction estimates from the results of the PlastiMouse.

So, in this study, a 2D-Net1 and 3D-Net2 model for denoising micro-CT images were trained on *ex vivo* mice scans and successfully evaluated on mice and phantom data. The computation time (with GPU) that the CNNs take to process (denoise) a typical scan is in accordance with the requirements of the preclinical field, where post-processing of a few minutes is an optimal case (here: sub-one minute processing was achieved). Important to also highlight is that the networks were obtained with training on a relatively limited number of datasets, which makes it easy to implement and translate the developed methods on other systems (and applicable to any modality).

4.1. Image Denoising Enhancement

Visual analyses show that both CNN approaches are more adaptive to suppress noise within targets of varying textural and structural details in the images compared to conventional filtering or iterative total variation image reconstruction techniques that are less adaptive. Measurements on the Molecubes QC CT phantom report a 9-fold increase in CNR for the method with the 2D-Net1 relative to LD image. For evaluations on the PlastiMouse, the 2D-Net1 has the lowest RMSE, highest PSNR and highest SSIM among comparator methods. Both the 2D-Net1 and 3D-Net2 also resulted in a good performance of CT number linearity for different contrast levels, suggesting that the CNN algorithms do not modify contrast patterns but rather improves contrast resolution. The results of the subtraction images are very encouraging as both CNN denoising solutions seem to do only minimal change to the real image content. Image quality assessment by MTF characterization further supports the fact that both methods (2D-Net1 and 3D-Net2) do not impact spatial resolution when denoising. Moreover, the 2D-Net1 denoised images are consistently favorably evaluated by the observers.

Visual results clearly showed that the 2D-Net1 and 3D-Net2 models did not learn to reproduce the noise present in the training target image but instead converged to predict a denoised image that is closer to a true, noise-free image. This behaviour of the networks being able to more easily pick up and learn the relevant image signal (rather than the noise content present in the target image) can be explained with the following argument. Two separate micro-CT acquisitions with different scan parameters (low and high dose settings) were performed to obtain input and target images that contain uncorrelated (independent) noise statistics. These two training image pairs (x_i and x_i') are considered to be identical up to their respective noise components (n_i and n_i') and share the same underlying

distribution of true signal content (which reflects the noise-free, or clean, image): $x_i = s_i + n_i$ and $x_i' = s_i + n_i'$. During model training, the network attempts to map a noisy input to a noisy target, and in the process of optimizing a loss function, the solution converges to the ground truth signal. The reasoning relies on a simple statistical argument: the estimate remains unchanged (i.e. equals the clean signal) as targets are replaced with random numbers whose expectations match the targets (noisy input-output). Details on the mathematical proofs are provided in [37,41,42]. The estimate of a ground truth image for the PlastiMouse (obtained from 10x HD scans averaged: Figure S-12, see Supplementary_Material.doc) provides support that both CNN denoising approaches indeed create an image that resembles a pseudo-ground truth.

4.2. Performance Comparisons between CNN Models and Associated Limitations

To explain why the 2D-Net1 has an improved denoising performance compared to the 3D-Net2, it should first be noted that CNN image restoration methods are often inherently less efficient in recovering various structural information in images due to the non-uniformity of noise property distribution and the mixture of texture appearances in CT images. Especially when pixel (voxel) wise loss functions (like mean-squared-error or mean-absolute-error) are implemented, the network tends to overlook perceptual effects for preserving the structures. For the 2D-Net1 denoising algorithm, the aforementioned limitation is addressed by training the network on whole image slices (instead of smaller image patches) and by imposing a 3-channel input layer. The latter implies that the network is presented with three adjacent CT slices, from which it creates 64 feature maps, thereby allowing to capture some coordinating dependencies across adjacent slices. Furthermore, the 2D-Net1 is trained on 12,000 slices in total (400 slices per mouse with 30 mice in the training set). Let us consider the example of one mouse dataset, for which 400 slices, each with an image size of more-or-less 200x200 pixels, are passed through the 2D-Net1. This implies that a total of 16,000,000 values (rough estimate) are used during training of 2D-Net1. In comparison, the 3D-Net2 is trained on a patch-by-patch basis with one channel at the input layer and uses 100 random patches of 32x32x32 in size per mouse for each training epoch (which equates to 3,000 patch units in total, for all 30 mice used in training – or differently said – equates to 3,276,800 voxels per mouse). These numbers clearly establish that the 2D-Net1 method has 4.8 times more parameter updates than the 3D-Net2, and this large difference in amount of training data can explain why the 3D-Net2 scores worse in (denoising) performance compared to 2D-Net1. Nonetheless, keeping in mind that the 2D-Net1 consists of 12,260,353 trainable parameters which is about 5.5 times more parameters than the 3D-Net2 with 2,213,377 trainable parameters, networks that have more learnable parameters generally require more data to achieve similar performance. Yet, while it could be argued that the number of learnable parameters

associated to each CNN method relatively corresponds to their difference in training data, the receptive field size of both networks impose different limitations on final performance. The receptive field of the models is 101^2 (or 101^3 for 3D case) as derived from the U-Net architecture employed for both CNNs (with three downsampling layers and convolutional layers of kernel size 3). While the 2D-Net1 (that is trained on the full image slices) can utilize its large receptive field, the 3D-Net2 trained on much smaller patches is hugely limited in its potential performance. It should thus highly be noted that the different training protocols employed for the two CNN methods (2D-Net1 and 3D-Net2) do not allow for a valid comparability between both approaches, and as such the presented results should not serve to generalize that a 2D CNN performs better than a 3D CNN for image denoising.

2D-Net2: 2D U-Net trained with patch-based approach

With the aim to present a more valid comparison, the 2D U-Net was re-trained with similar sized patches and on the same number of training samples as the 3D-Net2 method. In this regard, 100 random patches of 32x32 pixels in size per mouse (with 30 mice in the training dataset) were selected for each training epoch to update the 2D-Net2. Model training for the 2D-Net2 was completed in 29 epochs and took 14 minutes. Visual and quantitative results are presented in Figure 8. The 2D-Net2 and 3D-Net2 methods (both trained on a patch-based approach) showed similar remaining noise texture in the CNN-enhanced images with the 3D model offering a slight smoother appearance (e.g. inside the liver). The image quality metrics also reflect that 2D-Net2 and 3D-Net2 result in similar performance.

Taking into account the receptive field size of the U-Net models, the 2D-Net1 method achieves the most optimal results as it is able to utilize a larger region size in the input to associate an output feature, whereas the patch-based training approach with a much smaller patch size constrains how much information could be extracted to generate an associated feature. Thus, ideally, if not limited by computational power and memory usage, one would opt to use the full 3D dataset as input.

4.3. Potential for Dose Reduction using Deep Learning

The challenge remains to aim for the highest possible SNR with the most accurate spatial localization and temporal resolution, while controlling the amount of radiation given to the animal and not compromising image quality [3,4]. Most studies on DL-based denoising for low dose CT solely evaluate noise reduction performances and do not provide estimations on how much the dose could be reduced by. In comparison, we propose to derive dose reduction estimates from quantitative results and an

observer study. The quantitative results of the PlastiMouse phantom (i.e. 10x HR image, representative of a very high dose scan) and the assessments from the second observer study provide a starting point to discuss the potential of using the developed DL techniques to reduce X-ray radiation dose in preclinical settings.

Good agreement was found between both methodologies. Both CNN denoising methods were shown to achieve a 2-4x dose reduction, and quantitative measurements suggested that the 2D-Net1 could lead to a dose reduction factor of 3.2x (equivalent to a 68.8% reduction). Given the complex interplays between lower energy X-ray physics and biology, dose management (ALARA) should carefully consider the trade-off between image quality and biological response to radiation dose when designing micro-CT studies [43]. Our results are relevant for addressing the concerns on the radiation hazards of longitudinal experiments and/or multi-modality set-ups (e.g. micro-PET/CT).

4.4. Limitations and Generalizability of the CNNs

Quantitative measurements were conducted using the Molecubes QC CT phantom, however such a phantom may be too simplistic to realistically capture the image quality improvement. Phantoms are often designed to be used for quality control of the scanner's performance, but may not ideally be suited to reflect the performance of AI-based methods aimed at improving image quality. The PlastiMouse and a human observer study were thus included as additional assessment methods. It is well known that subjective evaluations are prone to observer bias, but they are clinically more relevant than phantom studies. The main shortcoming (but in some sense also strength) of our observer studies is that they were not based on a real (specific) diagnostic task, but instead asked for an overall image quality assessment.

Although the investigated DL approaches show promising image enhancement performance, a few limitations are worth noting. All the CNN methods were developed and tested with image datasets acquired on the same micro-CT scanner (X-CUBE) at only one X-ray tube potential (50 kV). Furthermore, training was solely performed on *ex vivo* micro-CT scans of mice obtained with a single noise level and reconstructed iteratively at 200 µm as input image. However *in vivo* imaging of rodents is typically conducted in preclinical practise which may suffer from slight image quality loss due to breathing and/or heart beating movements. Figure S-15 (see Supplementary_Material.doc) shows an example of an image acquired with the GP protocol (i.e. low dose image) from an *in vivo* study (ECD 21/63) that has been denoised using the 2D-Net1 model. It evidently shows that overall image quality in terms of noise reduction and contrast enhancement is remarkably improved. Moreover, the subtraction images mostly contain the noise residual and only present subtle changes.

(losses) in structural details at certain tissue boundaries. These results support the robustness of the network's denoising capabilities and promotes its applicability for *in vivo* micro-CT studies.

In addition, the applicability of the investigated CNN models to improve image quality on phantom data supports the robustness of its denoising performance (see Figure S-11 and S-12 in Supplementary_Material.doc). The 2D-Net1 method was also shown to still be competitive in handling cases of other noisy data (see Figure S-16 and S-17). Likewise, the 2D-Net1 was applied to a LD image reconstructed at 100 µm voxel size (see Figure S-18) to evaluate the applicability to higher resolution datasets, but to achieve best performance and improve generalizability, a more rigorous network optimization is needed whereby training is carried out with mixed input dataset (different noise levels). Besides aiming for improved noise reduction performance at various dose levels, the future step towards an even more enhanced image quality might be to consider a joint DL framework of denoising and super-resolution (train a network to transform low resolution, low dose to high resolution, high dose images).

5. Conclusions

This study has demonstrated the potential and feasibility of CNN-based denoising methods to predict (restore) high quality micro-CT images from noisy images acquired at reduced X-ray radiation doses. The 2D-Net1 and 3D-Net2 outperformed other more standard denoising methods: they were able to effectively suppress noise, while preserving fine structures and recovering contrast details without leading to blurry effects. The image quality scoring by 23 medical imaging experts also clearly indicated superior ranking for both CNN algorithms. Therefore, using such DL approaches allows to considerably lower radiation dose (with an estimated dose reduction factor of 3.2 for the 2D-Net1) which minimizes associated radiation hazards towards the laboratory animals. In the context of preclinical research, this offers promising future options for managing the cumulative severity effects of radiation in longitudinal studies, while offering an improvement in image noise and quality.

Acknowledgments: The authors thank dr. Bert Vandeghinste (Molecubes NV) for his technical support with the X-CUBE scanner, and all observers who participated in the study. The authors also thank dr. Brent van der Heyden and prof. Joel Karp for providing feedback.

Funding: This work was conducted in the scope of a master's thesis. Funding support is not applicable for this study.

Conflict of Interest: The authors declare that they have no conflict of interest.

References

1. Cunha L, Horvath I, Ferreira S, et al. Preclinical Imaging: an Essential Ally in Modern Biosciences. *Molecular Diagnosis & Therapy.* 2014;18(2):153-173.

2. Russell WM, Burch RL. *The principles of humane experimental technique.* London: Methuen & CO LTD; 1959.

3. Vanhove C, Bankstahl JP, Krämer DS, Visser E, Belcari N, Vandenberghe S. Accurate molecular imaging of small animals taking into account animal models, handling, anaesthesia, quality control and imaging system performance. *EJNMMI Phys.* 2015;2(31).

4. McCollough CH, Bushberg JT, Fletcher JG, Eckel LJ. Answers to Common Questions About the Use and Safety of CT Scans. *Mayo Clin Proc.* 2015;90(10):1380-1392.

5. Molinos C, Sasser T, Salmon P, et al. Low-Dose Imaging in a New Preclinical Total-Body PET/CT Scanner. *Front Med (Lausanne).* 2019;6:88.

6. Taschereau R, Chatziioannou AF. Monte Carlo simulations of absorbed dose in a mouse phantom from 18-fluorine compounds. *Med Phys.* 2007;34(3):1026-1036.

7. Wang J, Lu H, Li T, Liang Z. Sinogram noise reduction for low-dose CT by statistics-based nonlinear filters. Paper presented at: SPIE Medical Imaging2005.

8. Liu J, Ma J, Zhang Y, et al. Discriminative Feature Representation to Improve Projection Data Inconsistency for Low Dose CT Imaging. *IEEE transactions on medical imaging.* 2017;36.

9. Geyer LL, Schoepf UJ, Meinel FG, et al. State of the Art: Iterative CT Reconstruction Techniques. *Radiology.* 2015;276(2):339-357.

10. Willemink MJ, Leiner T, de Jong PA, et al. Iterative reconstruction techniques for computed tomography part 2: initial results in dose reduction and image quality. *Eur Radiol.* 2013;23(6):1632-1642.

11. Diwakar M, Kumar M. A review on CT image noise and its denoising. *Biomedical Signal Processing and Control.* 2018;42:73-88.

12. Thanh D, Prasath S, Le Minh H. A Review on CT and X-Ray Images Denoising Methods. *Informatica.* 2019;43:151-159.

13. Decuyper M, Maebe J, Van Holen R, Vandenberghe S. Artificial intelligence with deep learning in nuclear medicine and radiology. *EJNMMI Physics.* 2021;8(1):81.

14. Kim B, Han M, Shim H, Baek J. A performance comparison of convolutional neural network-based image denoising methods: The effect of loss functions on low-dose CT images. *Med Phys.* 2019;46(9):3906-3923.

15. Kulathilake KASH, Abdullah NA, Sabri AQM, Lai KW. A review on Deep Learning approaches for low-dose Computed Tomography restoration. *Complex & Intelligent Systems.* 2021.

16. Immonen E, Wong J, Nieminen M, et al. The use of deep learning towards dose optimization in low-dose computed tomography: A scoping review. *Radiography.* 2022;28(1):208-214.

17. Jing J, Xia W, Hou M, et al. Training low dose CT denoising network without high quality reference data. *Phys Med Biol.* 2022;67(8).

18. Bai T, Wang B, Nguyen D, Jiang S. Probabilistic self-learning framework for low-dose CT denoising. *Med Phys.* 2021;48(5):2258-2270.

19. Ulyanov D, Vedaldi A, Lempitsky V. Deep Image Prior. *International Journal of Computer Vision.* 2020;128.

20. Chen H, Zhang Y, Zhang W, et al. Low-dose CT denoising with convolutional neural network. Paper presented at: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017); 18-21 April 2017, 2017.

21. Chen H, Zhang Y, Kalra MK, et al. Low-Dose CT With a Residual Encoder-Decoder Convolutional Neural Network. *IEEE Trans Med Imaging.* 2017;36(12):2524-2535.

22. Kang E, Min J, Ye JC. A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction. *Med Phys.* 2017;44(10):e360-e375.

23. Yang W, Zhang H, Yang J, et al. Improving Low-Dose CT Image Using Residual Convolutional Network. *IEEE Access.* 2017;5:24698-24705.

24. Wolterink JM, Leiner T, Viergever MA, Isgum I. Generative Adversarial Networks for Noise Reduction in Low-Dose CT. *IEEE Trans Med Imaging.* 2017;36(12):2536-2545.

25. Yang Q, Yan P, Zhang Y, et al. Low-Dose CT Image Denoising Using a Generative Adversarial Network With Wasserstein Distance and Perceptual Loss. *IEEE Transactions on Medical Imaging.* 2017;PP.

26. Tang C, Li J, Wang L, et al. Unpaired Low-Dose CT Denoising Network Based on Cycle-Consistent Generative Adversarial Network with Prior Image Information. *Computational and Mathematical Methods in Medicine.* 2019;2019:1-11.

27. Kang E, Koo HJ, Yang DH, Seo JB, Ye JC. Cycle-consistent adversarial denoising network for multiphase coronary CT angiography. *Med Phys.* 2019;46(2):550-562.

28. Yin Z, Xia K, He Z, Zhang J, Wang S, Zu B. Unpaired Image Denoising via Wasserstein GAN in Low-Dose CT Image with Multi-Perceptual Loss and Fidelity Loss. *Symmetry.* 2021;13(1):126.

29. Han Y, Wu D, Kim K, Li Q. End-to-end deep learning for interior tomography with low-dose x-ray CT. *Physics in Medicine & Biology.* 2022;67(11):115001.

30. Akagi M, Nakamura Y, Higaki T, et al. Deep learning reconstruction improves image quality of abdominal ultra-high-resolution CT. *Eur Radiol.* 2019;29(11):6163-6171.

31. Chen L, Zheng L, Lian M, Luo S. A C-GAN denoising algorithm in Projection Domain for Micro-CT. *Molecular & Cellular Biomechanics.* 2019;17:1-8.

32. Yao W, Chen L, Wu H, Zhao Q, Luo S. Micro-CT image denoising with an asymmetric perceptual convolutional network. *Physics in Medicine & Biology.* 2021.

33. Clark D, Badea C. *Convolutional regularization methods for 4D, x-ray CT reconstruction.* Vol 10948: SPIE; 2019.

34. Muller FM, Vanhove C, Vandeghinste B, Vandenberghe S. Performance evaluation of a micro-CT system for laboratory animal imaging with iterative reconstruction capabilities. *Med Phys.* 2022.

35. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Paper presented at: MICCAI2015.

36. Kingma D, Ba J. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations.* 2014.

37. Lehtinen J, Munkberg J, Hasselgren J, et al. Noise2Noise: Learning Image Restoration without Clean Data. 2018.

38. Vandeghinste B, Vandenberghe S, Vanhove C, Staelens S, Van Holen R. Low-Dose Micro-CT Imaging for Vascular Segmentation and Analysis Using Sparse-View Acquisitions. *PLOS ONE.* 2013;8(7):e68449.

39. Fleiss JL, Cohen J. The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educational and Psychological Measurement.* 1973;33:613 - 619.

40. Fan L, Zhang F, Fan H, Zhang C. Brief review of image denoising techniques. *Visual Computing for Industry, Biomedicine, and Art.* 2019;2(1):7.

41. Wang C, Principe J. Training neural networks with additive noise in the desired signal. *Neural Networks, IEEE Transactions on.* 1999;10:1511-1517.

42. Wu D, Gong K, Kim K, Li Q. Consensus Neural Network for Medical Imaging Denoising with Only Noisy Training Samples. *ArXiv.* 2019;abs/1906.03639.

43. Meganck JA, Liu B. Dosimetry in Micro-computed Tomography: a Review of the Measurement Methods, Impacts, and Characterization of the Quantum GX Imaging System. *Molecular Imaging and Biology.* 2017;19(4):499-511.

Table 1: Scan parameter settings for two micro-CT protocols predefined on the X-CUBE: General-Purpose (GP) and High-Resolution (HR). Note that the X-ray tube remains switched on during the total scan time, so is not switched off between projections nor for detector read-out. The total radiation dose delivered to the animal can be estimated from the product of tube current and total scan time. Time per projection refers to the detector exposure time, when X-rays are captured. The imaging dose can be inferred from the product of tube current, number of projections and time per projection.

	General-Purpose (GP)	**High-Resolution (HR)**
Scan mode	Continuous	Continuous
Tube voltage (kVp)	50	50
Cathode tube current (μA)	75	350
Scan time in one bed position (s)	60	120
Dose given to the animal (mGy)	**4**	**42**
Number of projections	480	960
Time per projection (ms)	85	32
Imaging dose (mGy)	**3.06**	**10.75**

Table 2: Quantitative evaluations on the PlastiMouse, where RMSE, PSNR and SSIM are calculated with the ground truth, and on the Molecubes CT QC phantom, where image noise is measured. Results of the LD and HD image used as training input and target are included for reference.

	Noise (HU)	RMSE	PSNR	SSIM
Low dose	186.12 ± 19.42	0.113	39.83	0.974
Gaussian	103.79 ± 6.95	0.092	41.54	0.984
Median	48.95 ± 2.33	0.135	38.25	0.970
Wiener	47.61 ± 7.38	0.095	41.27	0.984
ISRA-TV	16.16 ± 0.85	0.110	40.03	0.977
2D-Net1	24.85 ± 4.33	0.080	42.75	0.989
3D-Net2	31.47 ± 1.31	0.087	42.05	0.985
High dose	81.29 ± 9.20	0.077	43.11	0.990

Table 3: Resultant estimates (derived from the quantitative analysis in Figure 6) for the dose reduction factors are presented as mean value (± standard deviation) calculated on the RMSE, PSNR and SSIM.

	Gaussian	Wiener	ISRA-TV	2D-Net1	3D-Net2	High Dose
Estimate of dose reduction	1.8 ± 0.1	1.7 ± 0.2	1.2 ± 0.1	3.2 ± 0.5	2.1 ± 0.2	3.9 ± 0.4

Figure Legends:

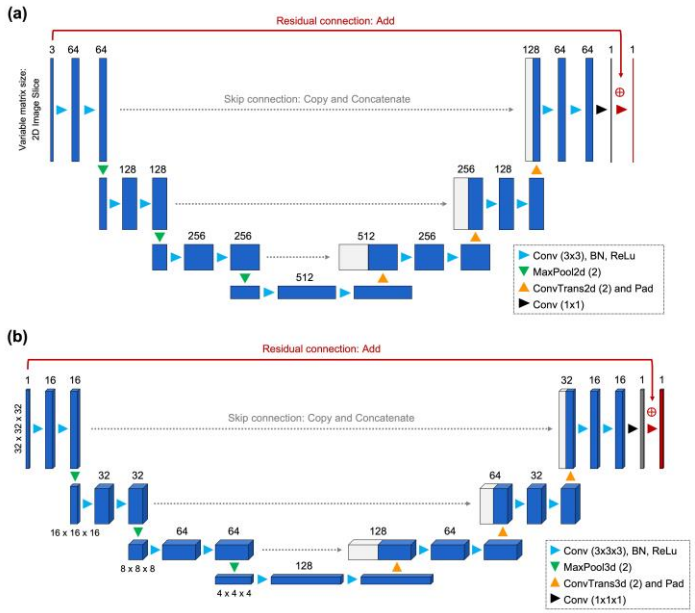

Figure 1: U-Net structures of the CNN denoising algorithms. **(a)** 2D-Net1 has 18 convolutional layers with 12,260,353 trainable parameters. Note that the model expects 3 image slices to be loaded as input (3-channel at the input). **(b)** 3D-Net2 has 18 convolutional layers with 2,213,377 trainable parameters. In both cases, the expansion layer input is padded for concatenation with the contraction layer output.

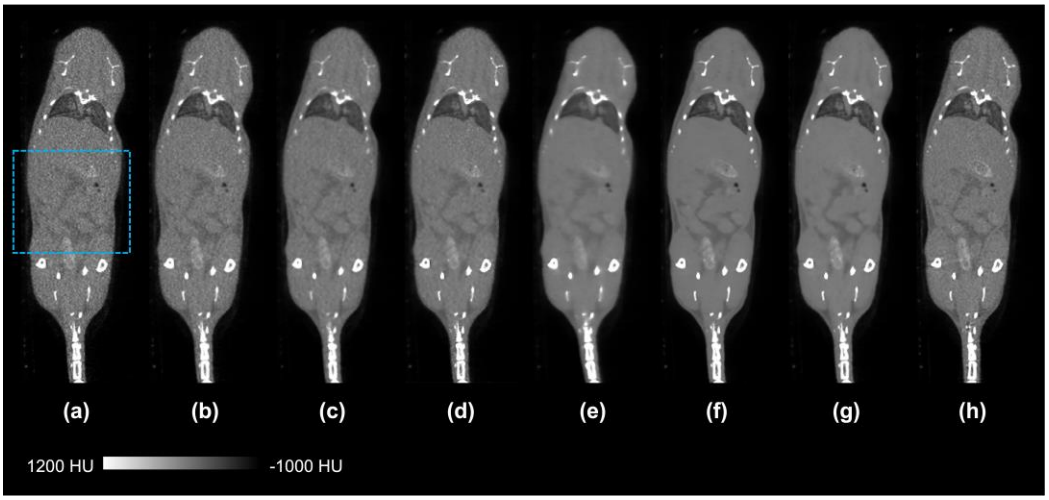

Figure 2: Denoising results from the ex vivo micro-CT mice scans. Coronal slice shown for comparisons: **(a)** LD, **(b)** Gaussian, **(c)** Median, **(d)** Wiener, **(e)** ISRA-TV, **(f)** 2D-Net1 denoised, **(g)** 3D-Net2 denoised, **(h)** HD image.

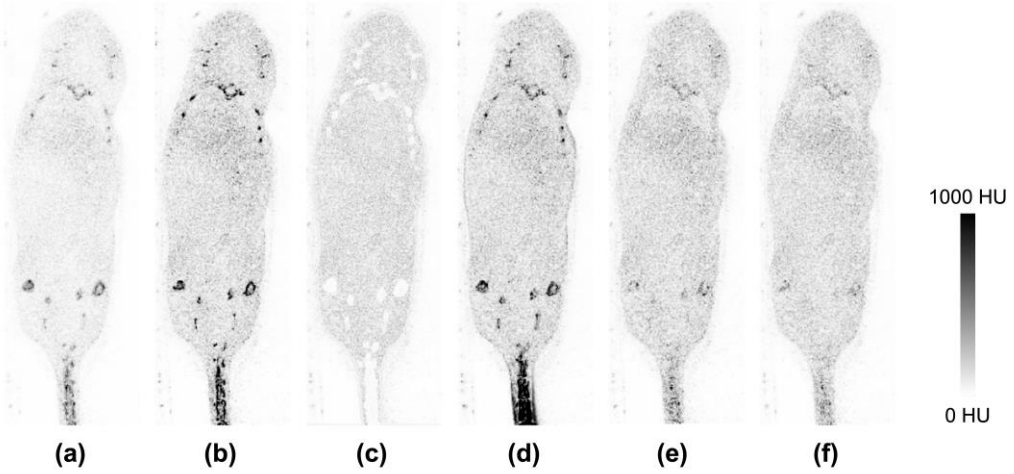

Figure 3: Subtraction images obtained by taking the absolute difference between the LD and the denoised images. The example shown here is for the image slices in Figure 2. Subtraction images with respect to the LD are compared between: (a) Gaussian, (b) Median, (c) Wiener, (d) ISRA-TV, (e) 2D-Net1, (f) 3D-Net2.

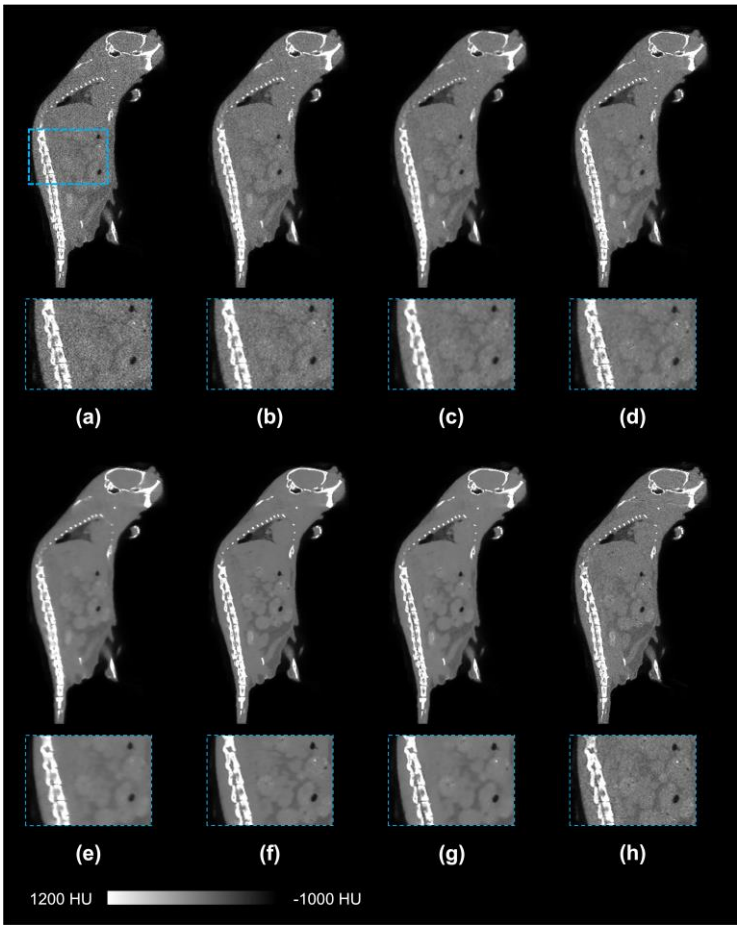

Figure 4: Denoising results from the ex vivo micro-CT mice scans. Sagittal slice shown for comparisons: (a) LD, (b) Gaussian, (c) Median, (d) Wiener, (e) ISRA-TV, (f) 2D-Net1 denoised, (g) 3D-Net2 denoised, (h) HD image. The dotted blue box zooms-in on a ROI as marked in (a).

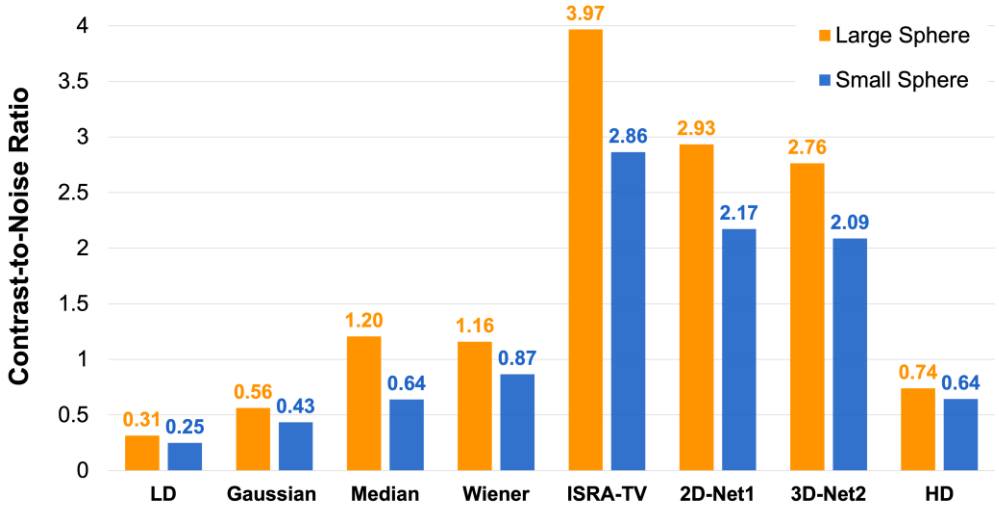

Figure 5: Contrast-to-noise ratio (CNR) measured in two contrast spheres (large: 10 mm diameter, small: 4 mm diameter) of the Molecubes QC CT phantom with respect to the background. CNR are compared between the different denoising techniques, and the CNR evaluated from the LD and HD image are also reported.

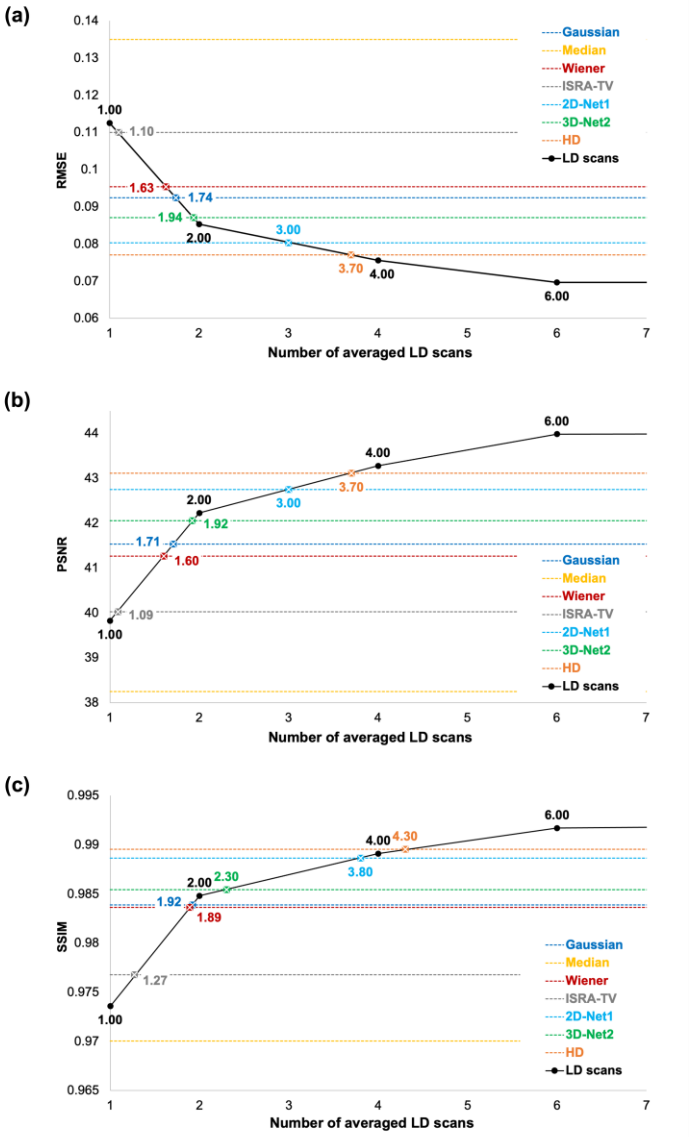

Figure 6: Quantitative analysis showing the image quality metrics as a function of the number of LD scans averaged to simulate images acquired at different dose levels (here depicted by the black scatter points at 1x LD, 2x LD, 4x LD and 6xLD). **(a)** RMSE, **(b)** PSNR, **(c)** SSIM. Note that the results of the Median filter are excluded, since it was found to have worse performance than the LD case.

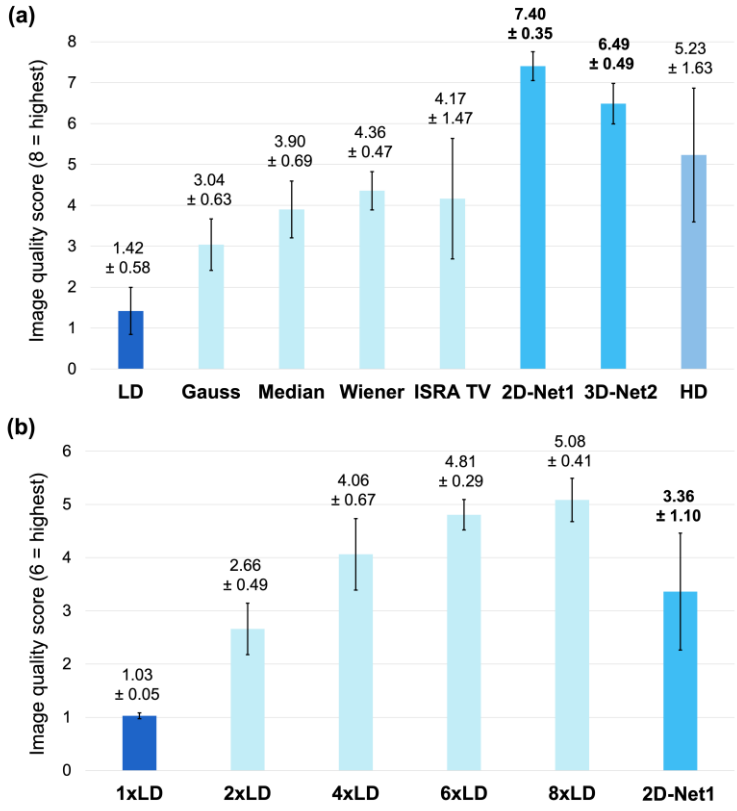

Figure 7: Results of the **(a)** first (n=23) and **(b)** second (n=18) observer study. Medical imaging experts were asked to rank different variants of micro-CT images according to the overall image quality using a Likert point scale (with 1 the poorest quality). Average score across all image sets from all observers ± standard deviation is reported.

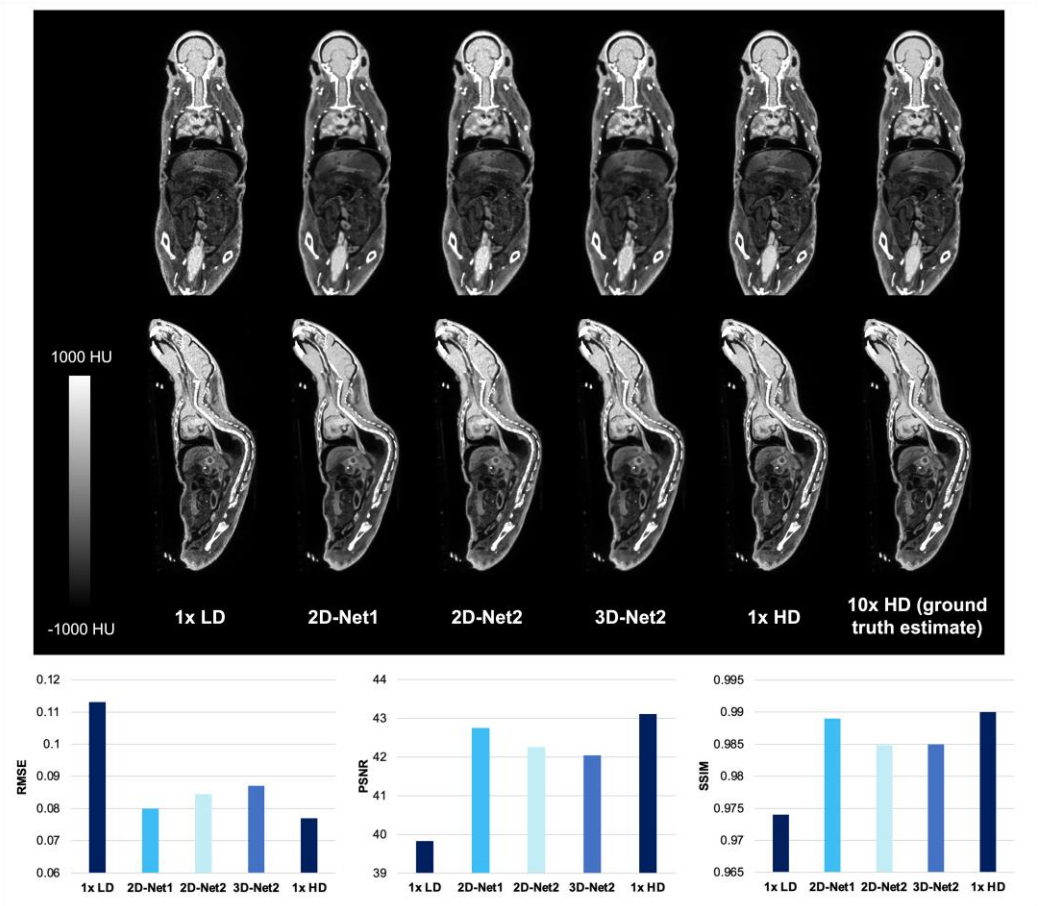

Figure 8: The three CNN-based denoising results from the PlastiMouse are compared: 2D-Net1 (2D U-Net trained on the full image slice), 2D-Net2 (2D U-Net trained on patches of 32x32) and 3D-Net2 (3D U-Net trained on patches of 32x32x32). The low dose (1x LD), high dose (1x HD) and ground-truth estimate (averaged 10x HD) image are also shown. The bar graphs report the evaluations on the PlastiMouse where RMSE, PSNR and SSIM are calculated relative to the 10x HD image.