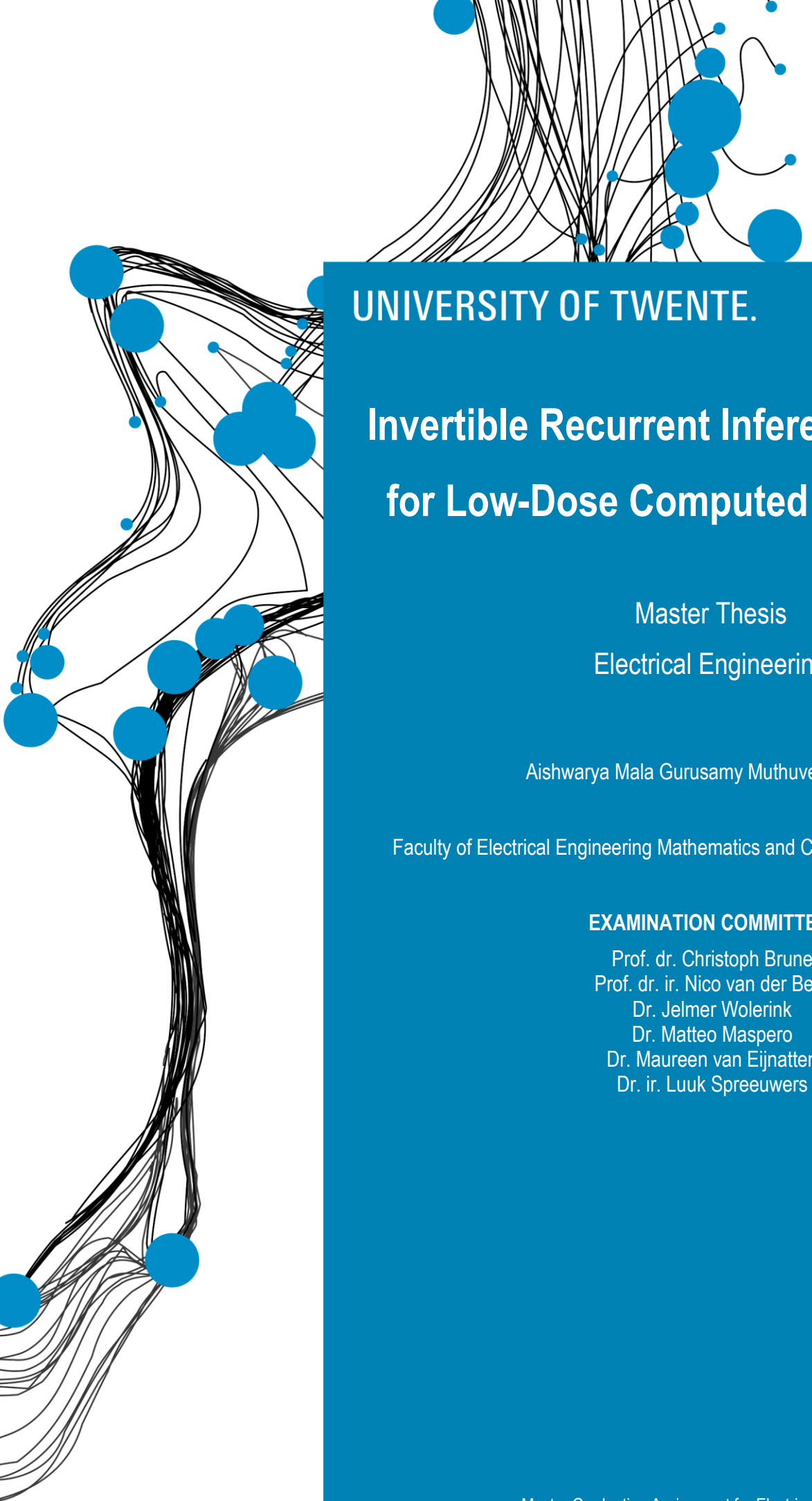

UNIVERSITY OF TWENTE.

Invertible Recurrent Inference Machines for Low-Dose Computed Tomography

Master Thesis

Electrical Engineering

Aishwarya Mala Gurusamy Muthuvelrabindran

Faculty of Electrical Engineering Mathematics and Computer Science (EEMCS)

EXAMINATION COMMITTEE:

Prof. dr. Christoph Brune
Prof. dr. ir. Nico van der Berg
Dr. Jelmer Wolerink
Dr. Matteo Maspero
Dr. Maureen van Eijnatten
Dr. ir. Luuk Spreeuwers

Master Graduation Assignment for Electrical Engineering
Enschede, 8th July 2021.

Invertible Recurrent Inference Machines for Low-Dose Computed Tomography

Aishwarya Mala GM
a.m.gurusamymuthuvelrabindran@student.utwente.nl

***Abstract*— Computed tomography (CT) is one of the most essential tools in medical imaging and physicians obtain visual representations of a patient's anatomy with the help of this x-ray based imaging modality. The carcinogenic properties of x-rays motivate the development of techniques that reduce radiation exposure. Low-dose CT is one such approach that uses low-intensity x-rays and a shorter exposure period to lower x-ray exposure. Due to reduced radiation dose, low-dose CT produces noisy projection data which in turn lowers the quality of the reconstructions. We turn to deep learning approaches to build quality reconstructions from noisy low-dose CT data. In this work, following its success in the fastMRI challenge that focuses on the reconstruction of undersampled magnetic resonance imaging (MRI) data, we investigate the use of invertible recurrent inference machines (iRIM) for low dose CT.**

Three iRIM models with likelihood gradient definitions of varying complexities were designed and trained uniformly on the LoDoPaB dataset. The image gradient iRIM model had the least complex likelihood gradient definition followed by the adjoint and FBP gradient iRIM models. Our adjoint gradient iRIM model performed the best and obtained an SSIM and PSNR mean and standard deviation of 0.8541 ± 0.1394 and 35.93 ± 4.74 dB on the LoDoPaB test data. It also achieved the best SSIM average of 0.8692 and secured an overall 4th position in the LoDoPaB challenge.

Furthermore, the generalization capabilities of the developed iRIM models were tested on three chosen categories – anatomy, low-dose simulation noise level and x-ray source beam geometry. Amongst the iRIM models the FBP gradient iRIM model proved to be the most capable on this front. The iRIM models were able to produce good results on the generalization capability tests but the performance degraded when the model had to handle high level noise contaminations. In conclusion, the iRIM framework proved to be suitable for low-dose CT but there are still a few scopes of improvement that could be considered to further enhance its robustness.

***Index Terms*— Computed tomography, deep learning, invertible neural networks, low-dose CT reconstruction, recurrent inference machine.**

I. INTRODUCTION

COMPUTED tomography (CT) is an x-ray based imaging modality that can be used to precisely visualize the internal structures of an object. A CT scanner consists of a source and detector pair that measures x-ray absorptions at different angles by moving around a target that needs to be scanned. These measurements are combined through reconstruction algorithms to obtain the internal volumetric density distribution of the target [1].

Since its advent, CT has been one of the most important tools in medical imaging. CT is predominantly used in preventive medicine, orthopaedics, and dentistry to detect pathologies such as infarctions, blood clots, calcifications and haemorrhages [2]. In oncology patients, CT can reveal the presence of tumours and help accurately locate them and estimate their size and structure. Due to this property, CT as a part of image-guided radiotherapy (IGRT) is used for treatment planning and further CT is also generally used to monitor the effects of cancer treatments on tumours [2].

The critical role of CT in clinical practice is evident from data in [3] that report the number of CT scans recorded in the member states of the European Union. The statistics presented show that in the five years leading up to 2018, there is a reported increase in the number of CT scans in all the member states of the European Union and these numbers are only expected to keep growing. This surge in the number of CT scans raises serious concerns about radiation-related cancer risks, as the ionizing radiation emitted by CT scanners can create free radicals or molecules causing damages to the tissues. Usually, the body can repair such damages, but when it does not, it can lead to the development of cancer [4,5]. Thus, to subdue the potential cancer risks linked to CT, there emerges a very compelling need to reduce the radiation dose but without making any compromises on the quality of the reconstructions that will be obtained for it.

One way to lower x-ray exposure is to decrease the x-ray tube current and shortened the x-ray exposure time. While this low-dose CT approach reduces the associated cancer risks, it produces noisy projection data due to reduced photon count. Thus, when analytical algorithms like filtered back projection (FBP) are employed for reconstruction, these noisy CT measurements can introduce unwanted image artifacts that degrade the quality of the reconstructions [6,7].

Iterative approaches to CT reconstruction provide a partial solution to the low-dose reconstruction problem. Here, the reconstruction is considered as an optimization problem and the algorithm tries to fit a solution iteratively. Various iterative CT reconstruction algorithms have proven to perform well on low-dose CT data at the cost of high computation time and complexity [6]. However, it might not always be feasible to wait for the algorithm to converge to obtain the reconstructions in practice. This led to a search for CT reconstruction approaches that rapidly produce results of high quality.

Deep learning (DL) has revolutionized many fields of study including medical imaging [8]. A highly successful application of DL in the field of medical imaging is image reconstruction. DL models outperformed traditional reconstruction approaches

in various imaging modalities like magnetic resonance (MRI), positron emission tomography (PET) and CT [8]. The incorporation of DL into low-dose CT reconstruction is thus a promising approach to overcome the shortcomings of the classical algorithms with improved reconstruction speed along with artifact reduction [9,10].

This work focuses on the invertible recurrent inference machines (iRIM), a deep learning model that was introduced by Putzky et al. [11,12]. Iterative inference problems can be unrolled in time and interpreted as a recurrent neural network (RNN). The recurrent inference machine (RIM) is one such RNN framework. Its primary advantage is that it does not need an explicit prior or inference procedure definition as these both will be implicitly learnt through its model parameters. The model was designed to learn an inference algorithm based on the given data and task. The invertible recurrent inference machine is an extension of the RIM that allows invertible learning during model training. This helps overcome memory constraints while also ensuring stable training even in the case of deep networks and large datasets. This is demonstrated by its performance on the fastMRI challenge, where the model was trained to reconstruct undersampled single and multi-coiled MRI data of the knee and brain to obtain results that are on par with the corresponding fully sampled data reconstructions [13,14].

Following the notable results produced by iRIM frameworks in undersampled MRI reconstruction, this work examines whether the benefits of the iRIM translate to the low-dose CT reconstruction problem. We aim at designing and training an iRIM model to perform low-dose CT reconstructions to yield results that match the quality of full-dose reconstructions. In addition, we also aim at testing the generalization abilities of the trained iRIM models across CT data collected over different anatomies, low dose simulations at different noise levels and different x-ray source geometries. Ultimately, at the end of this research we aim at understanding whether iRIMs can be developed into robust low-dose CT reconstruction models.

II. Related Work

The related work section starts with an introduction to CT reconstruction as an inverse problem. It is followed by a brief discussion on the classical CT reconstruction algorithms and their shortcomings. Lastly, the final part of the section focuses on the various DL-based approaches that have been proposed for CT reconstruction.

A. Inverse Problem

The inverse problem in CT reconstruction is shown in Fig. 1. The projection data or sinogram obtained from the scanners is denoted by y, the cross-sectional image reconstruction is denoted by x and the forward operator associated with CT, the Radon transform, is denoted by $\mathcal{R}$. Given the tomographic image x, the projection data y can be obtained by using the forward Radon transform operator $\mathcal{R}$. However, the inverse of Radon transform $\mathcal{R}^{-1}$ is ill-posed and calculating the cross-sectional image x from the projection data y becomes an inverse imaging problem [15].

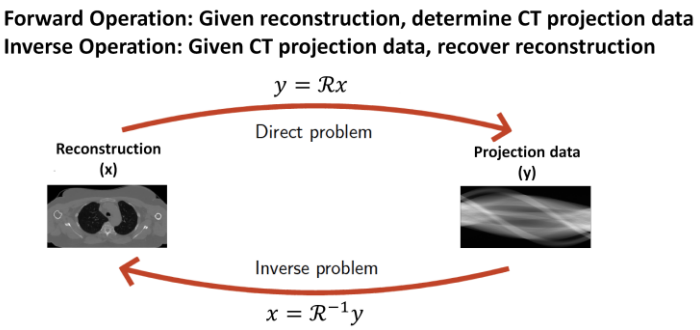

Fig. 1. CT Reconstruction – an inverse problem.

B. Classical CT Reconstruction Algorithms

As mentioned earlier, the inverse of the Radon transform $\mathcal{R}^{-1}$ is unbound and ill-posed [15]. There are several algorithms that are in use to recover the tomographic image from the projection data and they can be broadly categorized into analytical and iterative algorithms.

1) Analytical Algorithms

Fig. 2(a) depicts the working of a CT scanner with a parallel-beam x-ray source and Fig. 2(b) shows the sinogram obtained from the scanner data. The most commonly used analytical method to obtain the reconstruction from the sinogram is the filtered back projection (FBP) algorithm. As its name suggests, it consists of two steps, back projection (BP) and filtering.

BP is the adjoint of Radon transform. It is an operation that is based on the Fourier slice theorem and it smears the projection data obtained at every projection angle back into the image space. There is an over-weighing of low-frequency components during the back projection operation. Thus this step of the FBP algorithm is only capable of recovering a smoothened version of the cross-sectional image, as shown in Fig. 2(c). This problem is addressed by the filtering step. The deblurred result obtained after filtering can be seen in Fig. 2(d).

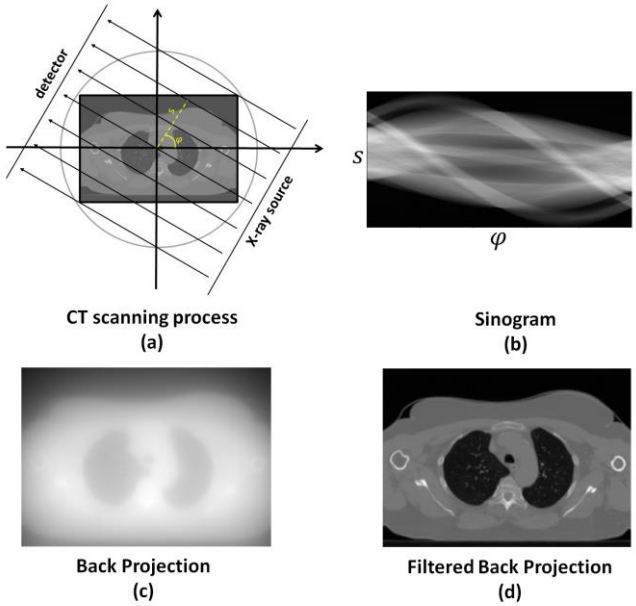

Fig. 2. (a) Parallel-beam CT scanning (b) The sinogram obtained after scanning (c) Result of back projection (d) Result of filtered back projection.

FBP is computationally very efficient and there are also variants of this algorithm to support different CT scanner geometries and acquisition techniques [6]. However, as a significant disadvantage of these analytical reconstruction algorithms, they demand noiseless data and regular projection angle distribution during data acquisition to yield images with acceptable quality. Thus, by imposing special conditions like low dose scanning in an attempt to reduce radiation exposure, it is made impossible to obtain good quality reconstructions using only analytical algorithms [6]. Iterative algorithms for CT reconstruction were developed to address this issue.

2) Iterative Algorithms

Iterative reconstruction (IR) algorithms treat CT reconstruction as an optimization problem and find a fitting solution iteratively. An ideal case of CT reconstruction can be denoted as a system of linear equations as shown in equation (1).

$$Ax = y \tag{1}$$

Here y is the projection data and x is the image to be reconstructed. Radon transform is discretized through the matrix A. Each row in A represents an x-ray beam and each column a pixel in the reconstructed image. Thus, each element a_{ij} in A represents the contribution of the i^{th} projection data to the j^{th} pixel. Rewriting equation (1) to solve for x and adding measurement noise n results in equation (2).

$$x = A^{-1}y + n \tag{2}$$

Ideally, a unique solution for x can be obtained if the matrix A and y are in the same column space, matrix A is invertible and all its columns are linearly independent (full column rank). However, in practice, in addition to the presence of noise, it is also not desirable to perform a CT scan with the required number of projection angles due to the need to reduce x-ray radiation exposure. Thus, the system of equations becomes ill-posed and an approximate solution is obtained using iterative methods. There are numerous IR algorithms available and one of the simplest known algorithms is the algebraic reconstruction technique (ART) [16]. It was introduced by Gordon et al. in 1970 and it was based on the Kaczmarz method that was initially proposed to solve a system of linear equations. ART and its variants come under a category of IR algorithms known as algebraic iterative algorithms.

Another class of IR algorithms is the statistical iterative algorithms that model CT reconstruction into probability models and arrive at the solution with the help of various statistical methods like maximum likelihood, least squares or maximum a posteriori estimation [17].

The results of iterative reconstruction algorithms are qualitatively better than those obtained using analytical reconstruction algorithms. Iterative reconstruction algorithms are equipped to produce good results even with the presence of noisy data and non-uniform projection angle distribution making them suitable for low-dose CT reconstruction. However, the image outputs might sometimes appear blotchy and as a major disadvantage, they are computationally costly and time-consuming [18]. DL algorithms were introduced into CT reconstruction to overcome such problems compromising the reconstruction quality.

C. Deep Learning in CT Reconstruction

Deep learning (DL) applications in CT reconstruction can be categorized into five major groups: image domain processing, data domain processing, a hybrid domain, deep reconstruction methods and iterative reconstruction combined with DL.

DL Models for Image Domain Processing: These methods take as input the image that was reconstructed using a traditional algorithm like FBP. The reconstructed image may contain noise and artifacts due to various reasons like dose limitations, inconsistencies during scanning or high attenuation materials such as metal implants inside the body during scanning. Image domain processing methods take the reconstructed image x, as input and try to compute a function $g(x) = \hat{x}$ such that $\hat{x}$ is as close to the ground truth (high-quality images without noise and artifacts) as possible. DL models can be trained by using a sufficient number of noisy CT reconstructions and the corresponding high-quality ground truths to achieve CT noise/artifact reduction.

Chen et al. in [19] have proposed a convolution neural network (CNN) model for low-dose CT denoising. The disadvantage of using CNN on medical images is that the structural details that are important for clinical diagnosis could get lost during the convolution and pooling operations. Thus, inspired by auto-encoders, deconvolution layers and shortcut connections were introduced into the CNN architecture. An improved model for CT denoising known as residual encoder decoder CNN (RED-CNN) is introduced in [20] with symmetrically placed convolution and deconvolution layers. Another CNN based denoising network for low dose CT is introduced by Kang et al. in [21]. The authors propose applying the deep-CNN model on the coefficients of wavelet transform. The aim is for the network to learn and remove the CT specific noise and artifact components present in the high-frequency components. The idea to use a generative adversarial network (GAN) for CT denoising was first introduced by Wolterink et al. in [22]. Following its success, there were various advancements proposed in that aspect of CT denoising too.

DL Models for Data Domain Processing: Following DL applications for image domain processing, another approach would be to focus on the data domain, that is, on the measurements obtained from the scanner. The lowering of x-ray beam radiation dose for low dose CT scanning results in noisy sinograms. If the noisy low dose sinogram data is represented by y, then the data domain processing methods aim at finding a function $f(y) = \hat{y}$ such that $\hat{y}$ is close to the ideal noiseless or full dose sinograms. The output sinograms are then reconstructed using conventional algorithms to obtained quality images.

Lossau et al. in [23] have proposed a data domain processing algorithm for metal artifact reduction in cardiac CT images. The process contains three CNNs. The first is a segmentationNet that takes as input the raw projection data and creates a metal shadow mask. The second network is an inpaintingNet that takes as input the metal shadow mask along with the original projection data to replace the metal-affected areas such that the reconstructions of the thus obtained sinograms are metal-free CT images. The third network, a reinsertionNet retrieves the

position of the metals based on a stack of partial back-projections and the metal shadow mask. Claus et al. in [24] have also attempted metal artifact reduction by using a three-layer fully connected network to correct the affected values in the sinogram. Although there are probabilities to obtained good results, data domain processing is very limited as the operations in the data domain have a tendency to easily produce image artifacts.

DL Models for Hybrid Dual-Domain Processing: Following a synergistic strategy, a dual-domain hybrid learning approach was introduced by combining the data-domain and image-domain learning methods. The noisy sinograms are processed by a network trained for data domain processing. The sinograms that were obtained as outputs after data domain processing are reconstructed using a simple reconstruction algorithm. The images thus obtained are fed into a network trained for CT image denoising.

Lee et al. in [25] have used a deep learning model that is based on a fully convolutional network and wavelet transform. The model was applied for sparse CT reconstruction through three different approaches – image domain, sinogram/data domain and hybrid domain. The results showed that the hybrid domain approach worked the best.

DL Models for Direct CT Reconstruction: A model under this category is trained to take as input the sinogram data and give as output the reconstructed image. Initially, efforts to use neural networks for CT reconstruction were based on the classical FBP algorithm. He et al. in [26] have proposed a framework known as iRadonMap where two networks are cascaded. The first network is divided into two segments. The first segment is a learnable, fully connected filtering layer. The second segment is the back-projection layer that transfers the filtered Radon projections into an image. The second network is a residual CNN that is used to refine the quality of the reconstruction. AUTOMAP is an image reconstruction DL framework proposed by Zhu et al. in [27]. It contains three fully connected layers that transform data from the sensor domain to the image domain. These are followed by two convolutional layers and a deconvolutional layer for noise and artifact reduction. The fully connected layers that are used for domain transform give this framework the capability to be generalized for different imaging modalities but at the cost of an increased number of parameters that need to be learned during training.

DL Models based on Iterative Reconstruction: Another DL approach to low-dose CT reconstruction would be to incorporate a DL model into an iterative reconstruction algorithm. Moraru in [28] has attempted to introduce a deep neural network into the simultaneous iterative reconstruction technique (SIRT). The network is trained to correct the update term at each iteration and has proved to facilitate faster algorithm convergence.

Most DL models discussed so far, unlike the traditional algorithms, do not work based on any mathematical derivations or the physical modelling of the imaging modality. This may become a weakness and a cause for sub-optimal results. To overcome this effect, DL-based reconstruction networks are developed in reference to conventional iterative reconstruction algorithms. One such DL framework is the learned primal-dual algorithm proposed by Adler and Öktem in [29]. The forward model is accounted into the DL model by unrolling a proximal primal-dual optimization method where the proximal operators are replaced by CNNs. The model produced notable results when trained for low-dose CT reconstruction while also proving to be appropriate for time-critical clinical applications.

The recurrent inference machine (RIM), introduced by Putzky et al. in [11], is a similar kind of DL model that learns an inference algorithm by unrolling it in time and interpreting it as a recurrent neural network (RNN). Invertible recurrent inference machine (iRIM), as its name suggests, is an invertible network implemented with RIM as the base. It was proposed by Putzky et al. in [12] with the aim to increase the expressiveness of a model while ensuring stable training with constant and reasonable memory requirements. Following its success in undersampled MRI reconstruction [14], we study its performance in low-dose CT reconstruction in this research work.

III. Method and Materials

This section begins with a description of the iRIM framework. Following that, the dataset that was used during model training and testing is described along with its pre-processing steps. Further, the definitions of the various image evaluation metrics that are later used during the model performance assessments are included.

A. Invertible Recurrent Inference Machine

Recurrent Inference Machines (RIM): To know about RIM, we start with an RNN. These are neural networks that have an internal memory. To compute the output at a given step, the RNN takes into consideration the current input and the output learnt from the previous input. This process is quite similar to an iterative algorithm that updates its current prediction based on its previous prediction. Thus an iterative inference algorithm can be unrolled and implemented as an RNN.

RIM is one such RNN framework. It depicts an inverse problem as a probability model and uses an RNN to implement a statistical iterative algorithm that can optimize towards a maximum a posteriori (MAP) solution. For an inverse problem that is defined as per equation (2), equation (3) below shows a MAP solution for x in terms of $p(y|x)$, a likelihood term representing the noisy model and $p_\theta(x)$ a parametric prior over x.

$$\max_x \log p(y|x) + \log p_\theta(x) \tag{3}$$

Equation (4) represents a recursive function in its simplest form that can be used to arrive at this MAP inference for x where γ_t is the step size or learning rate and x_t is the MAP estimate for x at iteration t.

$$x_{t+1} = x_t + \gamma_t \nabla(\log p(y|x) + \log p_\theta(x))\,(x_t) \tag{4}$$

The update term for x is modified and written in the form of a function with learnable parameters ϕ as shown in equation (5) below.

$$x_{t+1} = x_t + g_\phi(\nabla_{y|x}, x_t) \tag{5}$$

If we use $\nabla_{y|x}$ to denote $\nabla(\log p(y|x))\ (x_t)$ and ∇_x to denote $\nabla(\log p_\theta(x))\ (x_t)$, combining equations (4) and (5) will give the below definition for the function g_ϕ.

$$g_\phi(\nabla_{y|x}, x_t) = \gamma_t(\nabla_{y|x} + \nabla_x) \tag{6}$$

Since the update term includes only the gradient, there is no need to learn an actual prior for x. On that account, the modification done on the update term allows the model to directly learn the gradient function $\nabla_x = f(x_t)$ instead of learning a prior and then computing its gradient. Secondly, this modification also removes the need to estimate the step size γ_t by including it implicitly into the learnable parameters of the inference model along with the prior gradient function ∇_x. The likelihood model that is part of the input $\nabla_{y|x}$ is the source of task-specific information. Unlike the prior gradient function and step size, $\nabla_{y|x}$ is kept as a separate input. Due to this property, the likelihood model of a RIM framework can be modified without changing the learned parameters ϕ of the inference model. This gives a RIM framework the capability to generalize across related tasks after just suitably adapting the likelihood model.

Fig. 3 shows an outline of the update term modification. The red boxes represent the internal data-independent modules and the blue boxes represent the external data-dependent modules. Fig. 3(a) is a representation of the recursive function given by equation (4). Fig. 3(b) represents the modified model with trainable parameters that produces estimates based on the feedback from the likelihood model and that is given by equation (5).

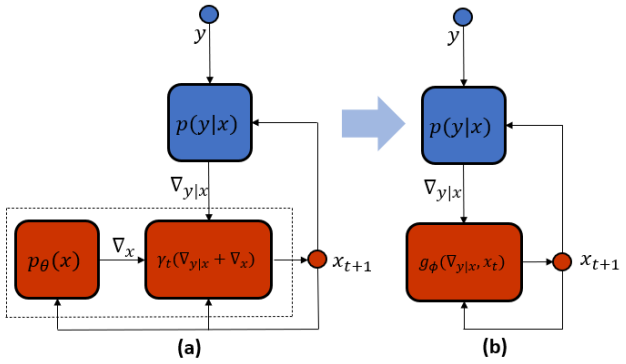

Fig. 3. (a) Block diagram representing the recurrent function for MAP inference. During an iteration t, the likelihood $p(y|x)$ and $p_\theta(x)$ are calculated using the current estimate x_t. These are used as input to calculate the update term to produce the new estimate x_{t+1} as shown in equation (4). (b) The modified model with the prior and step size estimation merged into one model with trainable parameters ϕ. Graphical representation of equation (5). Figure sourced from [11]

Equation (5) is fitted into an RNN that has a network h_ϕ within it to model the function g_ϕ and after adding a latent memory variable s_t, the update equations of the RIM take the below form.

$$x_{t+1}, s_{t+1} = x_t + h_\phi(\nabla_{y|x}, x_t, s_t) \tag{7}$$

A training loss function $\mathcal{L}$ is defined to measure the similarity between the model prediction at a given step x_t, a function of the model parameters ϕ and the corresponding ground truth x. During the back propagation steps in the training process, the model calculates a total loss $\mathcal{L}_{total}$ based on the loss obtained at each RNN step. Equation (8) shows the total loss for a RIM model with T steps. The model learns to improve the quality of the prediction through this loss function.

$$\mathcal{L}_{total} = \sum_{t=1}^{T} \mathcal{L}(x_t(\phi), x) \tag{8}$$

Fig. 4 shows an unrolled RIM and its correspondence to the update equation in (7) could be seen. The loss estimated at each step during training is denoted by dashed lines. The dashed lines also indicate that this process occurs only during model training when the ground truth is fed into the model.

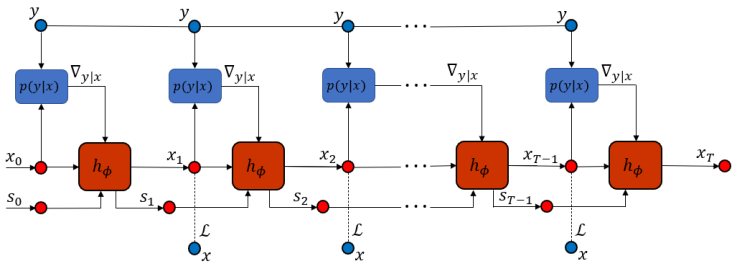

Fig. 4. An unrolled recurrent inference machine. The input and the likelihood gradient module are represented in blue showing that they are data-dependent and externally defined. The learning modules and all the associated machine states are shown in red showing that they are data-independent internal modules of the RIM model. Figure sourced from [11]

As mentioned earlier, the likelihood gradient module is an external module and needs to be defined explicitly according to the inverse problem at hand. Below are three possible likelihood gradient definitions for CT reconstruction that we explore in this study.

Image Gradient: This is a likelihood gradient that has the simplest definition computationally without involving any CT forward operators within it. As shown in equation (9), it is the difference between x_t, the model prediction at the given iteration t and the FBP reconstruction of the low-dose sinogram input y. The calculations are entirely done in the image domain and it does not involve the use of any information that is derived from the distribution of the model estimate in the sinogram domain. The aim of this likelihood gradient definition is to compare the performance of a model that is trained with a likelihood that does not include the CT forward operator against performances of models that are trained with likelihoods that include the CT forward operator.

$$\nabla_{y|x} = x_t - FBP(y) \tag{9}$$

Adjoint Gradient: The adjoint likelihood gradient is defined as per equation (10). The model estimate x_t, using the Radon transform $\mathcal{R}$, is projected into the sinogram domain, where its difference with the original low dose sinogram y is computed. The likelihood gradient is used to compute the update term for the image estimate and thus it needs to be brought back to the image domain from the sinogram domain. For this purpose, the

adjoint of Radon transform denoted by $\mathcal{R}^H$ is used as the back projection operator. This likelihood gradient definition is computationally not very costly and at the same time also includes information from the projection domain which is expected to improve the model reconstruction capability.

$$\nabla_{y|x} = \mathcal{R}^H(\mathcal{R}(x_t) - y) \tag{10}$$

FBP Gradient: The definition of this likelihood gradient is similar to the previously defined adjoint gradient but instead of using back projection as the inverse operator, it uses the filtered back projection as shown in equation (11). Back projection yields as output a blurred or smoothened version of the actual reconstruction due to over-weighing of the low-frequency components. Thus, an additional filtering step is added into the inverse operator to compensate for the over-weighing. This modification is expected to increase the computation complexity of the likelihood gradient and at its cost feed more relevant information into the model to push it towards better quality CT reconstructions.

$$\nabla_{y|x} = FBP(\mathcal{R}(x_t) - y) \tag{11}$$

Invertible Recurrent Inference Machines (iRIM): It is to be noted from the RIM architecture that h_ϕ is a network that is contained within each of the RNN steps. Therefore, training the RIM will include back propagation through the RNN and within that another back propagation to train the network in each RNN step. This is referred to as back-propagation through time and it forces the imposition of limits over the depth and complexity of the networks to keep the memory requirements of the whole model within check. To overcome this constraint, Putzky et al. in [12] introduced the invertible RIMs (iRIMs) by proposing to use the reversible neural network architecture inspired by the work of Gomez et al. in [30]. This allows the restoration of intermediate activations from post activations and in turn memory saving by removing the need to store these activations. Thus, iRIMs are architecturally modified RIMs that are invertible. Apart from memory saving the iRIM frameworks are also equipped to remove training instabilities and have the ability to accommodate large volumes of training data. For further information on how a RIM framework is modified into an iRIM please refer to appendix I. More information can also be obtained from [12, 13].

B. Dataset and Pre-processing Steps

The public low dose parallel beam (LoDoPaB) dataset [31] was used to train the iRIM model for low-dose CT reconstruction. It is a highly heterogeneous collection of thoracic scans with tube peak voltages ranging from 120 kV to 140 kV and tube currents ranging from 40 mA to 627 mA with a mean of 222.1 mA. This dataset contains 35820 training images from 632 patients, 3522 validation images and 3553 test images from 60 patients each. The projection data corresponding to each CT image is obtained using the Radon transform function provided by Python's operator discretization library (ODL) [32] under a chosen setup of 1000 projection angles and 513 detector elements paired with a parallel beam x-ray source. A low dose setting is modelled by using Poisson-distributed noise with a mean photon count of 4096 per detector element before attenuation.

The pre-processing steps to obtain the ground truth images and the steps that were used to simulate the low dose projection data are listed in algorithm 1. The pre-processing for ground truth images starts with im_HU, which is the available CT reconstruction in Hounsfield unit (HU) that is centre cropped to the size of 362 px × 362 px. Linear attenuations (μ) are computed from the HU values to obtain im_MU as shown in step 1. μ_{max} is the maximum linear attenuation value calculated based on the maximum HU value and im_MU is normalised using μ_{max} and clipped to obtain ground truth images with pixel values in the range [0,1].

For the low dose projection data simulation, im_HU is first dequantized by adding uniform noise from the interval [0,1). Following this, HU values are converted to linear attenuations and then normalized and clipped to obtain im_norm with pixel values in the range [0,1]. To avoid committing an inverse crime [33], before the forward projection operation, im_norm is upscaled from 362 px × 362 px to 1000 px × 1000 px using bilinear interpolation. The projection data corresponding to the upscaled image is computed using ODL's Radon transform function. Photon count per detector pixel is computed by assuming an initial photon count of 4096 and applying Beer-Lambert's law on the projection data as shown in step 6. Step 7 models the low dose setting with the help of Poisson noise and replaces any zero photon counts with 0.1 so that the log transform in the next step produces only finite valued outputs. Beer- Lambert's law is used again to acquire the low dose projection data and it is divided by μ_{max} to make it compatible with the previously obtained normalized ground truth images. More detailed explanations can be found in [30, 34].

Algorithm 1 Steps to obtain ground truth images and to simulate low dose projection data

$\mu_{air} = 0.02$, $\mu_{water} = 20$, $\mu_{max} = 81.36$

Pre-processing steps for ground truth images:

1. $\text{im\_MU} = \text{im\_HU} * \frac{\mu_{water} - \mu_{air}}{1000} + \mu_{water}$
2. $\text{GT} = \text{clip}(\text{im\_MU}/\mu_{max},\ \text{min} = 0; \text{max} = 1)$

Low dose projection data simulation steps:

1. $\text{im\_HU} \mathrel{+}= \text{dequantization\_noise} \sim \mathcal{U}(0,1)$
2. $\text{im\_MU} = \text{im\_HU} * \frac{\mu_{water} - \mu_{air}}{1000} + \mu_{water}$
3. $\text{im\_norm} = \text{clip}(\text{im\_MU}/\mu_{max},\ \text{min} = 0; \text{max} = 1)$
4. $\text{im\_upscaled} = \text{bilinear\_interpolation}(\text{im\_norm})$
5. $\text{proj\_data} = \text{Radon\_transform}(\text{im\_resize})$
6. $\text{photons} = \exp(-\text{proj\_data}) * 4096$
7. $\text{noisy\_photons} = \max(0.1, Poisson(\text{photons}))$
8. $\text{proj\_data}_{LD} = -\ln(\text{noisy\_photons}/4096)/\mu_{max}$

C. Evaluation Metrics

The dataset used contains ground truth images. Thus, all the evaluation metrics are full-reference objective image quality measures that compare original distortion less ground truth CT images with the test reconstructions. Below are the image similarity measures that were used to report the results obtained on the LoDoPaB dataset. We opted to use the same set of metrics for our experiments and performance evaluations.

Peak Signal to Noise Ratio (PSNR): PSNR is the ratio between the maximum possible image intensity and the noise. The maximum possible image intensity usually lies far away from the range of expected values in the case of CT images. Therefore, it is replaced by the difference between the highest and lowest intensity values in the ground truth image. Below is the equation for PSNR calculation with this modification for a test image $\hat{x}$ and its corresponding ground truth x both of size $M \times N$.

$$PSNR = 10\, log_{10}\left(\frac{[\max(x) - \min(x)]^2}{MSE(x, \hat{x})}\right) \tag{12}$$

$$MSE(x, \hat{x}) = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}[x(m,n) - \hat{x}(m,n)]^2 \tag{13}$$

Structural Similarity Index Measure (SSIM): SSIM is an assessment method that is based on the characteristics of the human visual system (HVS). SSIM works on computing the overall structural similarity between the ground truth and test images by comparing the normalized local patterns of pixel intensities with the help of a sliding window. The sliding window creates M local regions each on the ground truth and test image. Mean pixel intensities (μ), variance (σ) and covariance (Σ) are calculated on the M local regions on the ground truth and test images and are used to compute the SSIM value as per the below equation. C_1 and C_2 are constants that were used to stabilize the division. The window size was set to 7×7. $C_1 = 0.01 * \max(x)$ and $C_2 = 0.03 * \max(x)$. Python's scikit-image library was used for SSIM calculation [35].

$$SSIM = \frac{1}{M}\sum_{j=1}^{M}\frac{(2\hat{\mu}_j\mu_j + C_1)(2\Sigma_j + C_2)}{(\hat{\mu}_j^2 + \mu_j^2 + C_1)(\hat{\sigma}_j^2 + \sigma_j^2 + C_2)} \tag{14}$$

IV. Experiments and Results

A. iRIM Models Training and Testing

1) Model Training

For the iRIM model structure, a framework similar to the one that was used by Putzky et al. in [14] for the fastMRI challenge was chosen. The chosen iRIM framework contains an RNN with 8 steps and a network that is 12 layers deep within each step. The number of channels was set to 64. Weight sharing was set to false, which means the parameters were not shared across the RNN steps. The resulting iRIM model contains about 275M learnable parameters. The models were trained on the LoDoPaB dataset with a batch size of 8. Each batch is a set of 8 low-dose sinograms along with the respective ground truth images. The model was trained using a structural similarity loss function with Adam optimizer for 10 epochs where one epoch corresponded to the model looping through the 35820 samples in the LoDoPaB training data. The learning rate was set to 10^{-4} and was reduced by a factor of 10 after 5 epochs.

Three iRIM models were built, each with one of the likelihood gradient functions defined in section III.A. configured into it. Python's ODL library [32] was used to implement all the CT operators in the likelihood gradients. In order to make the operators compatible with the LoDoPaB dataset, all operators were initialised with a reconstruction geometry that corresponded to 1000 projection angles, 513 detector elements and a parallel beam x-ray source. The FBP operator, apart from the said configurations, needs a suitable filter. The ODL library uses the Ram-Lak filter by default and the same was used for the FBP operator in the likelihood gradient too. Apart from the likelihood gradient definitions, all other model configurations and training setups were kept the same for the three models. Each model takes as input the low dose sinogram y and the FBP reconstruction of the low dose sinogram. The low dose FBP reconstruction is used as the initial model estimate x_0 and the low dose sinogram y is used in likelihood gradient estimation.

The validation loss curves obtained during the training of the three models are shown in Fig. 5. The plot shows the average SSIM values obtained at the end of each validation cycle during which the models were evaluated on the 3522 images in the validation data. At the end of the training, for each likelihood definition the best model that gave the highest average SSIM against the validation data was saved. The three saved models were then evaluated on the test data.

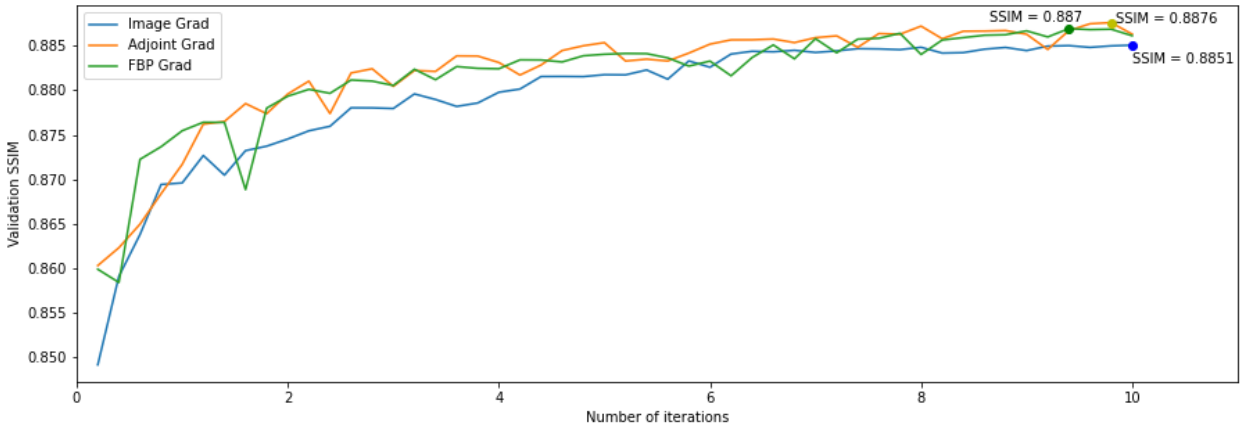

Fig. 5. Validation curves of the three iRIM models during training. The best models and the corresponding average SSIM value obtained on the validation data are also shown in the plot.

It is also important to note that the training time and the reconstruction time of the models increased with an increase in the likelihood gradient complexity. The FBP likelihood gradient definition is the most complex. The corresponding model's training time for one epoch is approximately 36.7 hours and the time required by the FBP gradient iRIM model for one reconstruction is close to 0.96 seconds on an NVIDIA Tesla-V100 GPU with 32 GB RAM. The next less complex model, the adjoint gradient iRIM model had a training time of 16.1 hours for an epoch and a reconstruction time of 0.45 seconds on the same GPU. The least complex model, the image gradient iRIM model had a training time of 16.8 hours and a reconstruction time of 0.44 seconds on an NVIDIA Tesla-P100 GPU with 16 GB RAM.

2) Results on the LoDoPaB Test Data

All results are compared to two baselines. First, the FBP reconstructions of the low-dose sinograms. Second, the outputs that were obtained using the reference U-Net that was provided along with the LoDoPaB dataset [36]. A U-Net is a convolution neural network that consists of an encoding and decoding path. The encoding path contains convolution layers that remove redundant spatial information and performs a function analogous to feature extraction. The decoding path contains a series of up-convolution layers and concatenations that build the output from the features and information obtained from the encoding path. The U-Net model in this case is used as a denoiser and the chosen architecture contains 10 layers with approximately 610K learnable parameters. It takes as input the FBP reconstruction of the low-dose sinogram and performs a denoising operation to remove all the unwanted streaks and image artifacts. The U-Net model was trained on the LoDoPaB dataset using mean squared error (MSE) loss and Adam optimizer for 20 epochs. More details on the U-Net model used can be found in [30, 35].

Box plots representing the model performances on the LoDoPaB test data can be found in Fig 6. In addition to the quartiles shown in these boxplots, we report the mean and standard deviations, as is commonly done for the LoDoPaB dataset. The iRIM model with adjoint gradient likelihood showed the best performance amongst all the models with a mean SSIM of 0.8541. The same is also reflected by its PSNR mean of 35.93 dB which is the highest amongst the models in comparison.

Fig. 7 shows a random sample ground truth image and the corresponding model outputs. On the second row are the error images that were obtained by calculating the difference between the ground truth and each of the model outputs. The model outputs visually do not appear significantly different but the SSIM and PSNR values can be seen to be varying. The error images also signify the fine differences amongst the model outputs

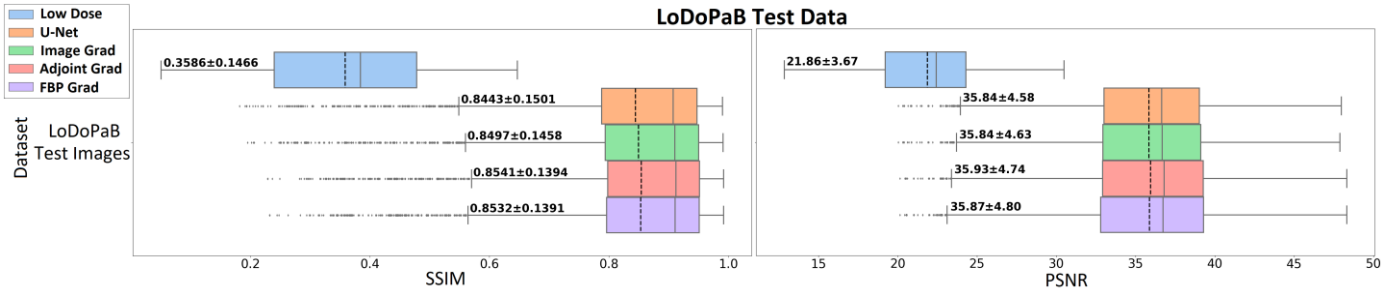

Fig. 6. Boxplots showing the distribution of the SSIM and PSNR values evaluated on the LoDoPaB test outputs obtained from the three iRIM models. The evaluations on the U-Net and low-dose FBP outputs are also shown for comparison. The corresponding mean (also represented by the dashed line) and standard deviation are displayed beside each plot.

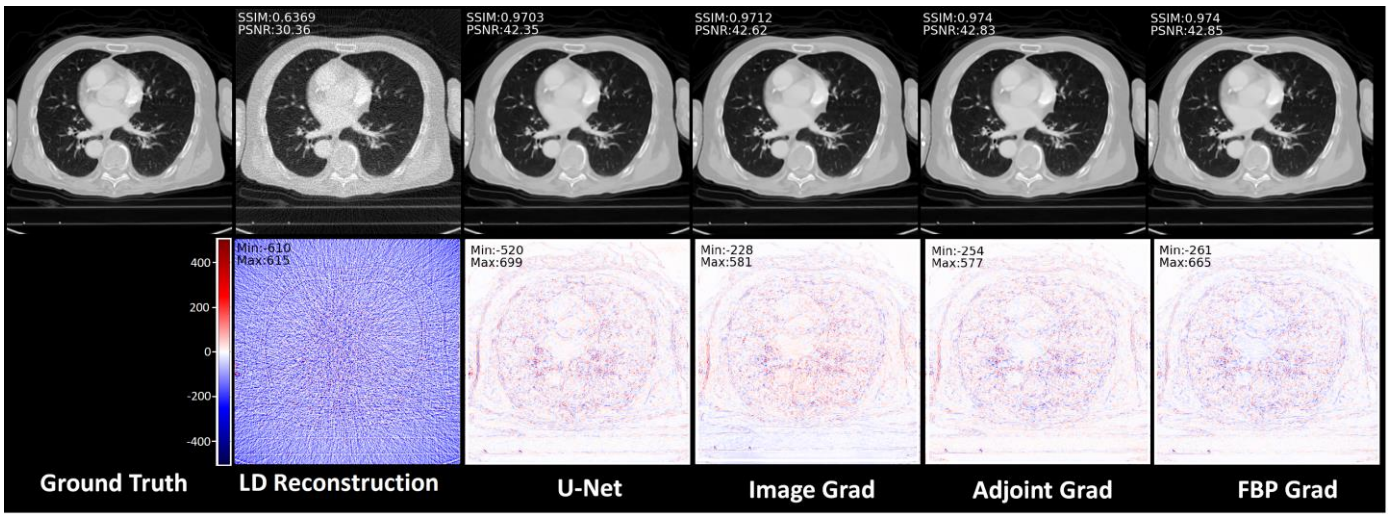

Fig. 7. Sample from LoDoPaB test outputs obtained from the three iRIM models, the U-Net and the FBP reconstruction of the low dose sinogram. The corresponding ground truth image is also provided for comparison. Image window: [-1001, 424] HU. The respective error images that were obtained by calculating the difference between the ground truth and each of the model outputs are on the second row. Image window: [-500, 500] HU.

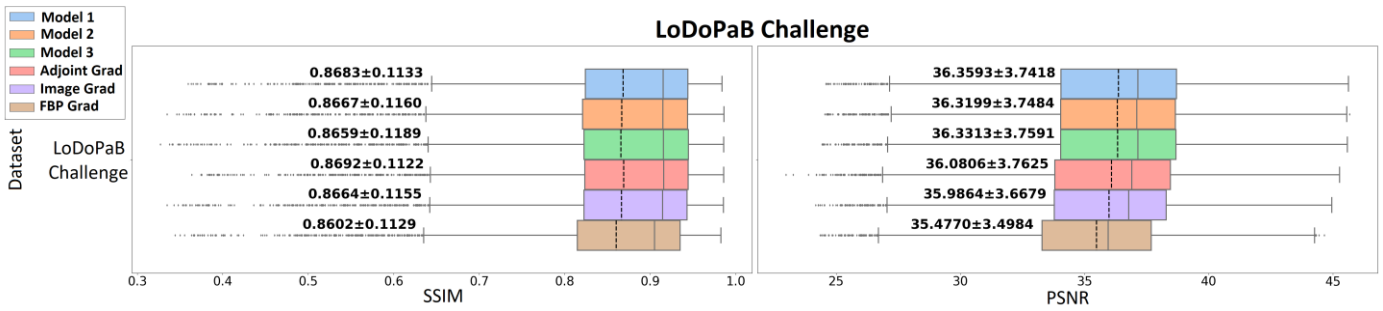

Fig. 8. Boxplots representing the performance of the iRIM models in the LoDoPaB challenge. Model 1, 2 and 3 are the low-dose CT reconstruction algorithms that currently hold the first, second and third positions respectively in the challenge.

3) Results of the LoDoPaB Challenge

To compare our models with the other existing low dose reconstruction algorithms, we evaluated them on the LoDoPaB challenge dataset and submitted the results. This is a separate dataset for which we did not have access to the ground truth reconstructions. The adjoint gradient iRIM model was able to achieve an SSIM mean of 0.8692 and standard deviation of 0.1122 and a PSNR mean of 36.08 dB and a standard deviation of 3.76 dB. Our model achieved the best SSIM average among all the submissions and secured an overall 4th position. The FBP gradient adjoint model yielded an average SSIM of 0.8602 and a standard deviation of 0.1129. The image gradient iRIM model gave an average SSIM of 0.8664 and a standard deviation of 0.1155 on the LoDoPaB challenge dataset. Fig. 8 contains boxplots to compare the performances of the three iRIM models against the low dose CT reconstruction models that currently hold the top three positions in the LoDoPaB challenge. Model 1, 2 and 3 represent the low dose CT reconstruction algorithms that currently hold the first, second and third position in the LodoPaB challenge. The boxplots show that the iRIM models have the ability to produce high SSIM means but lag on the PSNR front. There are notably large differences in the mean PSNR values of the other low-dose CT reconstruction algorithms and the iRIM models. The iRIM models were trained using a SSIM based loss function and that might be a plausible reason behind this particular behaviour of the iRIM models.

B. Model Generalization Capability Tests

From the performance of the iRIM models on the LoDoPaB challenge, it could be inferred that the models have the capability to generalize well within the dataset. However, this is not sufficient to analyse if the model has the capability to perform adequately in a real-world setting. All three trained iRIM models were thus evaluated for their generalization properties outside the LoDoPaB dataset. Three focus areas were chosen and the generalization experiments along with the results obtained subsequently are discussed below. All the evaluations are done on the iRIM and U-Net models that were trained on the LoDoPaB dataset and there was no sort of retraining involved during the generalization experiments.

1) Anatomy

The distribution and spatial configuration of soft tissue, bones and other structures within the body vary significantly as one moves from the head to the toes. Yet, it is essential for a CT reconstruction model to be able to perform well on any CT data irrespective of the anatomical structure being scanned. Thus, the performances of the trained iRIM models on CT data that belong to anatomical regions of the body that are structurally different from the training data need to be analysed. The data in the LoDoPaB dataset belong to the thoracic region. Therefore, to check the generalization capabilities of the trained models over different anatomies, CT data that belonged to the pelvic, head and neck regions were selected. The 'other anatomy' dataset created contains a total of 3453 two-dimensional CT images of which 2211 are head & neck images collected from 12 different patients and the remaining 1242 are pelvic images collected from a set of 10 patients. The peak tube voltage was set at 120kV for all the scans and the tube current ranged from 100 mA to 290 mA for the pelvic scans and from 128 mA to 271 mA for the head and neck scans. The ground truth images and the low-dose projection data were obtained using the same set of LoDoPaB pre-processing steps as listed in Algorithm 1.

Fig. 9 summarizes the performances of the models on the other anatomy dataset. The results obtained on the head and neck CT images and on pelvic CT images are kept separate. From Fig. 9 it can be clearly inferred that the adjoint and the FBP gradient models were able to generalise better to different anatomies than the image gradient iRIM model. Out of the iRIM models, the FBP gradient iRIM model showed the best performance. On the other hand, the U-Net was also able to extend its performance on the other anatomy dataset and was even able to produce results that were better when compared to the results from the image grad iRIM model. Fig. 10 shows the model outputs of two randomly chosen inputs along with the corresponding ground truths and error images. The first sample is a head CT and the second sample is a pelvic CT. The error and output images that belong to the image gradient iRIM model stand out and additionally prove that the adjoint and FBP gradient iRIM models are able to generalize better anatomy-wise due to the added complexities in their likelihood gradients.

On the head and neck CT data, the adjoint and the FBP grad iRIM models were able to show performances that were better than the performance of the U-Net but only by a narrow margin. Mean SSIM calculated on the U-Net outputs was 0.9566 and thus the narrow margin might be due to already high SSIM values obtained on the U-Net outputs. On analysing the pelvic CT data results, the performance of the U-Net with a mean SSIM of 0.9507 was found to be better than all the iRIM models. Along with the boxplots in Fig. 9, this result is also clearly reflected by the pelvic output images in Fig. 10.

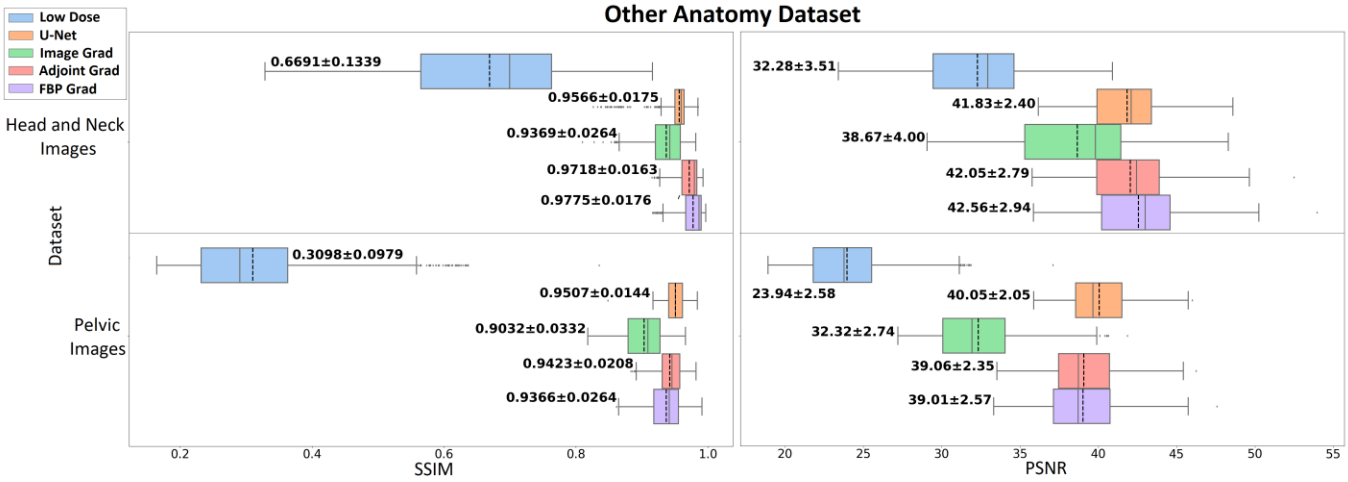

Fig. 9. Boxplots showing the distribution of SSIM and PSNR values calculated on the model outputs that correspond to the other anatomy dataset. The performances on the head and neck CT data and on the pelvic CT data are displayed separately.

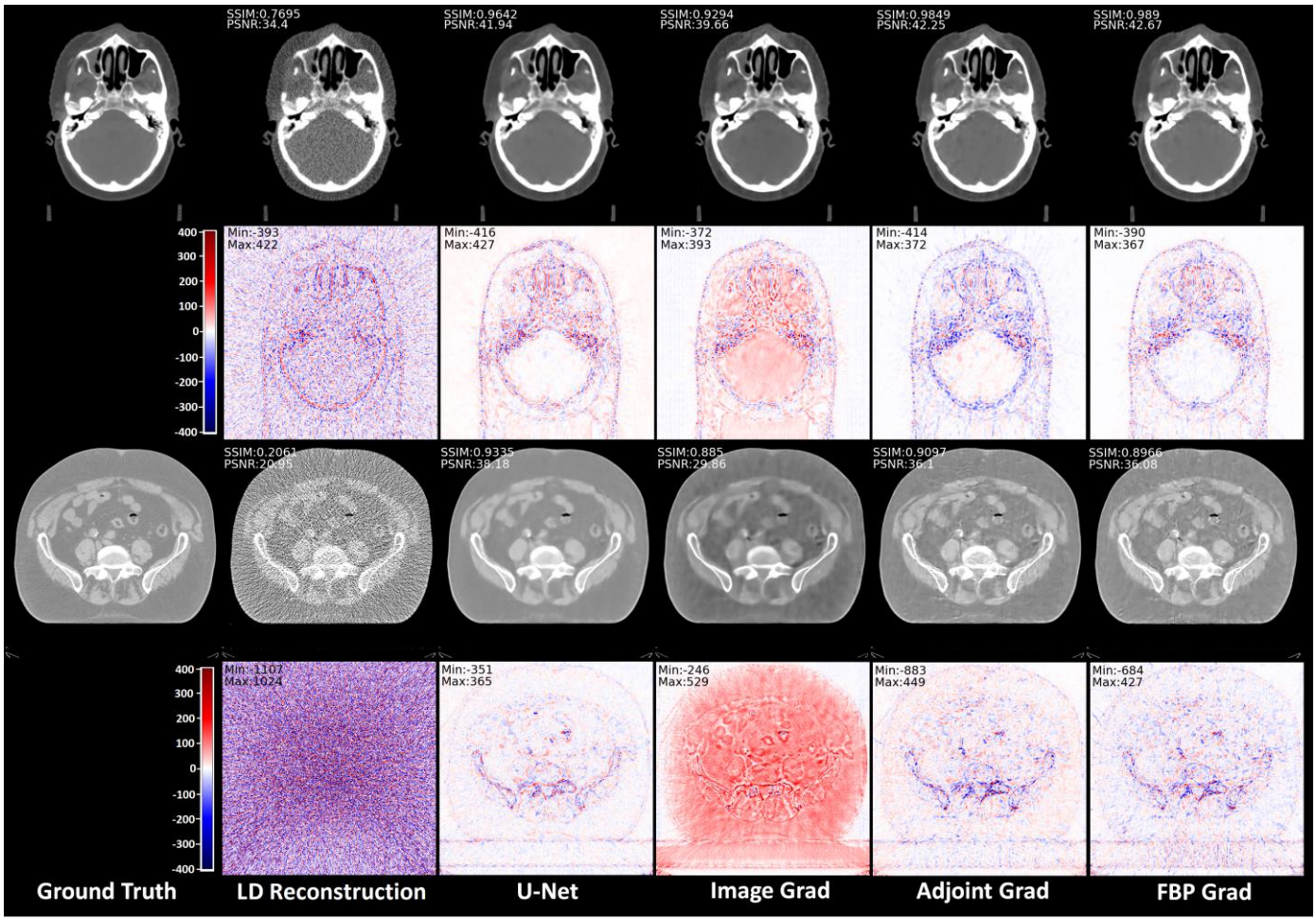

Fig. 10. Sample model output from other anatomy dataset along with their ground truth and error images. The first two rows correspond to head and neck CT data and the last two rows correspond to pelvic CT data. The head CT images are displayed in an image window of [-390, 628] HU. For pelvic CT images the window is approximately [-512, 221] HU. All the error images are in the window [-400, 400] HU.

2) Noise Level

The low dose projection data in the LoDoPaB dataset were simulated using Poisson noise. Although it is a good approximation, there are great chances for real-world low-dose conditions to be very different. Thus the performances of the models at different noise levels were examined to measure the immunity of the models against noise variations. For this purpose, two datasets were forged from all the 3553 test CT images of the LoDoPaB dataset and the 3453 other anatomy CT images. While the ground truth images remained the same, during the low dose sinogram simulations the mean photon count of the Poisson noise was reduced from 4096 to 2048. This

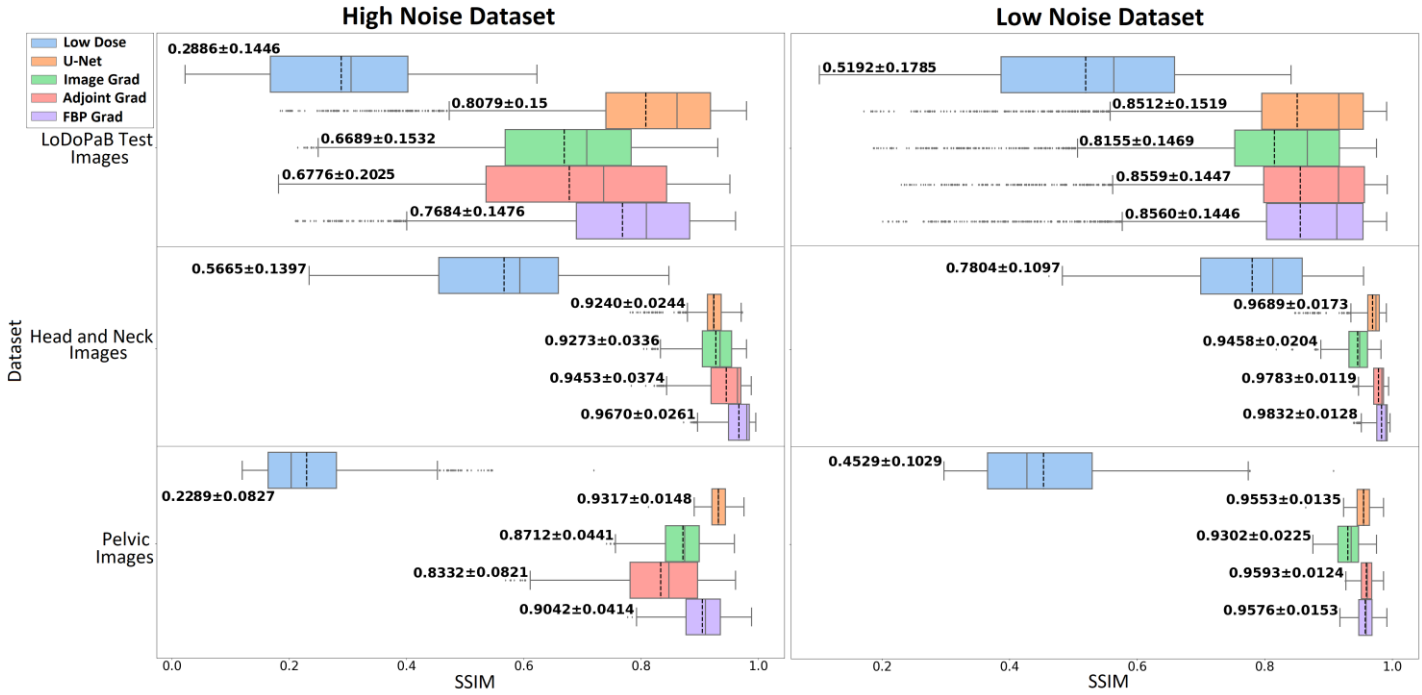

Fig. 11. Boxplots showing the distribution of the SSIM values calculated on the results obtained from the high and low noise dataset. The performance of the models on the LoDoPaB test images, head and neck CT images and on the pelvic CT images are displayed separately.

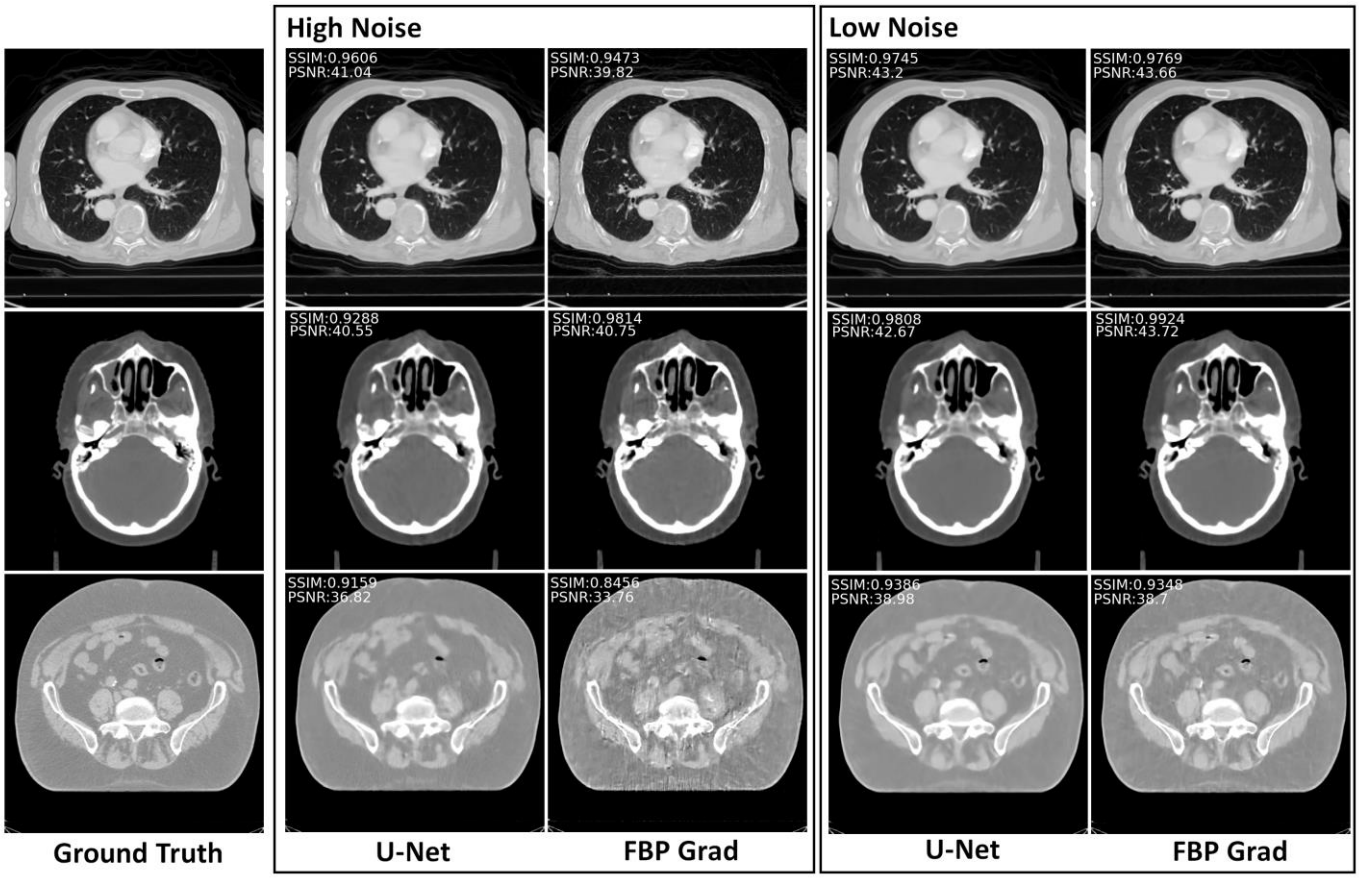

Fig. 12. Sample outputs and the corresponding ground truth images from the high and low noise dataset results obtained from the U-Net and the FBP gradient iRIM model. The same set of ground truths from Fig. 7 and Fig. 10 are used. The FBP gradient iRIM model performed the best out of the three iRIM models and thus only its outputs are displayed along with the corresponding U-Net outputs. The previously mentioned image windows are also maintained. On the first row for the LoDoPaB test images the display window is [-1001, 424] HU. For the head images in the second row it is [-390, 628] HU and for pelvic images in the last row it is [-512, 221] HU.

resulted in noisier projection data and these were used to create what was named the ‘high noise’ dataset. Similarly, low-dose sinogram simulations using Poisson noise with an increased mean photon count of 8192 that resulted in less noisy projection data were used to create the ‘low noise’ dataset.

The trained iRIM models and the reference U-Net were evaluated on both the high and low noise datasets. Fig. 11 contains the boxplots representing the distribution of the SSIM values. The performances of the model on different anatomies are displayed separately for better analysis. The boxplots show that the FBP likelihood gradient iRIM model was able to consistently maintain its performance and thus in Fig. 12 only the FBP gradient iRIM model output images are displayed along with the U-Net outputs for comparison.

The boxplots in Fig. 11 show that the iRIM model is able to hold its performance on the high noise head and neck images while losing it on the high noise pelvic and LoDoPaB test CT data. The FBP gradient iRIM model gave the highest SSIM average of 0.9670 on the head and neck images. On both the high noise pelvic and LoDoPaB test images, the U-Net was able to show better performance than all the iRIM models with the highest SSIM average of 0.9317 and 0.8079, respectively. However, following the reduced performance on the high noise dataset, the performances of the iRIM models on the low noise dataset were considerably better, especially on the pelvic images. From Fig. 12 it could be noted that the improvements on the pelvic outputs are even visually evident. The adjoint and FBP gradient iRIM models performed better than the image gradient model and the FBP gradient iRIM model shows a slight edge over the adjoint version. The adjoint and FBP likelihood gradient iRIM models were also able to outperform the U-Net on all three anatomies while the image gradient iRIM model lagged behind the U-Net on all the cases considered.

3) Source Beam Geometry

All the datasets that were used or that were created so far contained sinograms that were simulated under a chosen setup with a parallel beam x-ray source. In order to test the cross-task generalization capability of the iRIM models, a ‘low dose fan beam’ (LoDoFaB) dataset was created by replacing the parallel beam forward projection operator in the projection data simulation process with a fan beam forward projection operator.

Before the performance assessments of the iRIM models, the parallel beam CT operators within their likelihood gradient function definitions need to be replaced with their fan-beam counterparts. The suitably collected outcomes from both iRIM and U-Net models were examined to deduce appropriate inferences on the abilities of the iRIM models to generalize across tasks. Fig. 13 contains the corresponding boxplots and Fig. 14 contains the set of sample outputs and ground truths along with the error images. Again, out of the three iRIM models, the FBP gradient iRIM model had the best performance. Thus only its output is displayed in Fig. 14 along with the corresponding outputs from the U-Net for comparison.

Out of the three iRIM models, the FBP gradient iRIM model was found to have the best performance on the LoDoFaB dataset. While the image gradient iRIM model shows average performance, the adjoint gradient model breaks down. An issue of model incompatible fan beam adjoint operator is suspected to be the reason behind this observation. The results show slack in the U-Net performance especially in head and neck CT data while the FBP gradient iRIM model holds its performance. The error images that are displayed along with the model outputs in Fig. 14 also show that the iRIM model outputs are better than U-Net outputs. The iRIM model again shows performance issues on the pelvic images but that has been a common observation amongst all the previously obtained outputs. Thus, it could be seen that the iRIM model was able to retain its performance while the performance of the U-Net degraded.

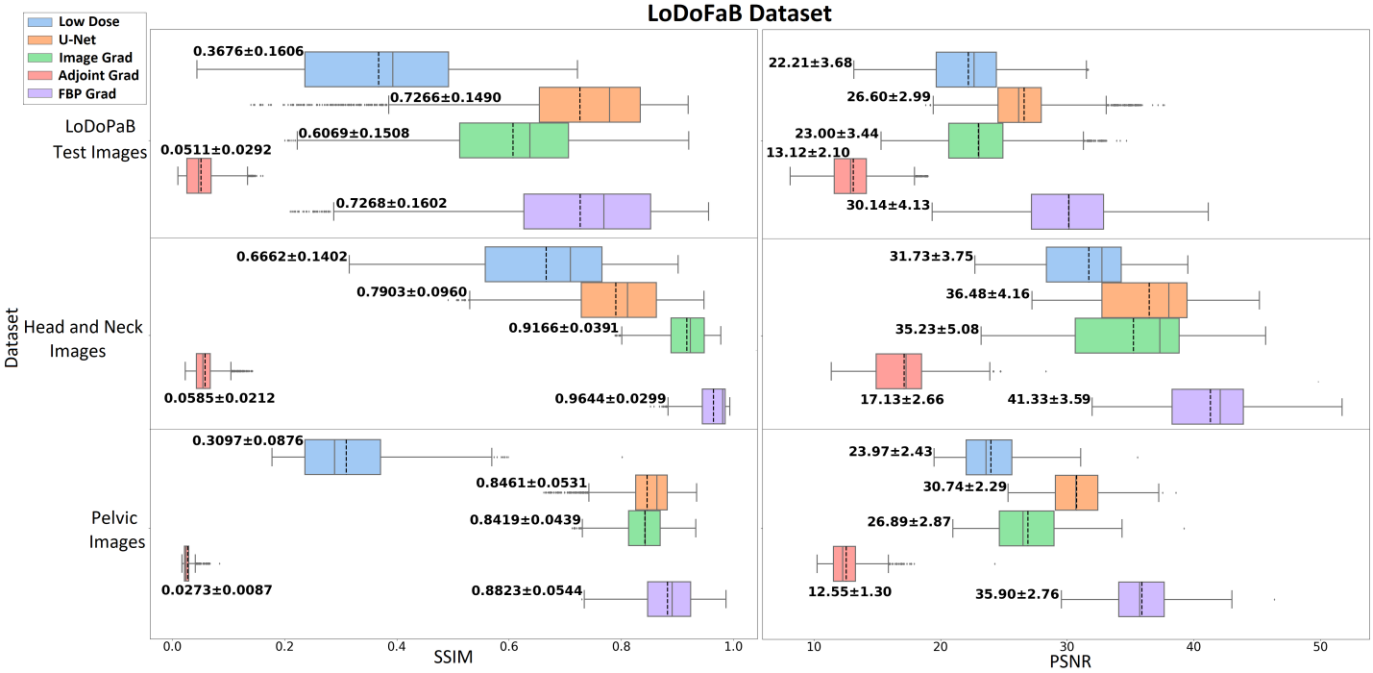

Fig. 13. Boxplots showing the distribution of the SSIM and PSNR values calculated on the results obtained from the LoDoFaB dataset. The performances of the models are categorised based on the anatomy of the CT data.

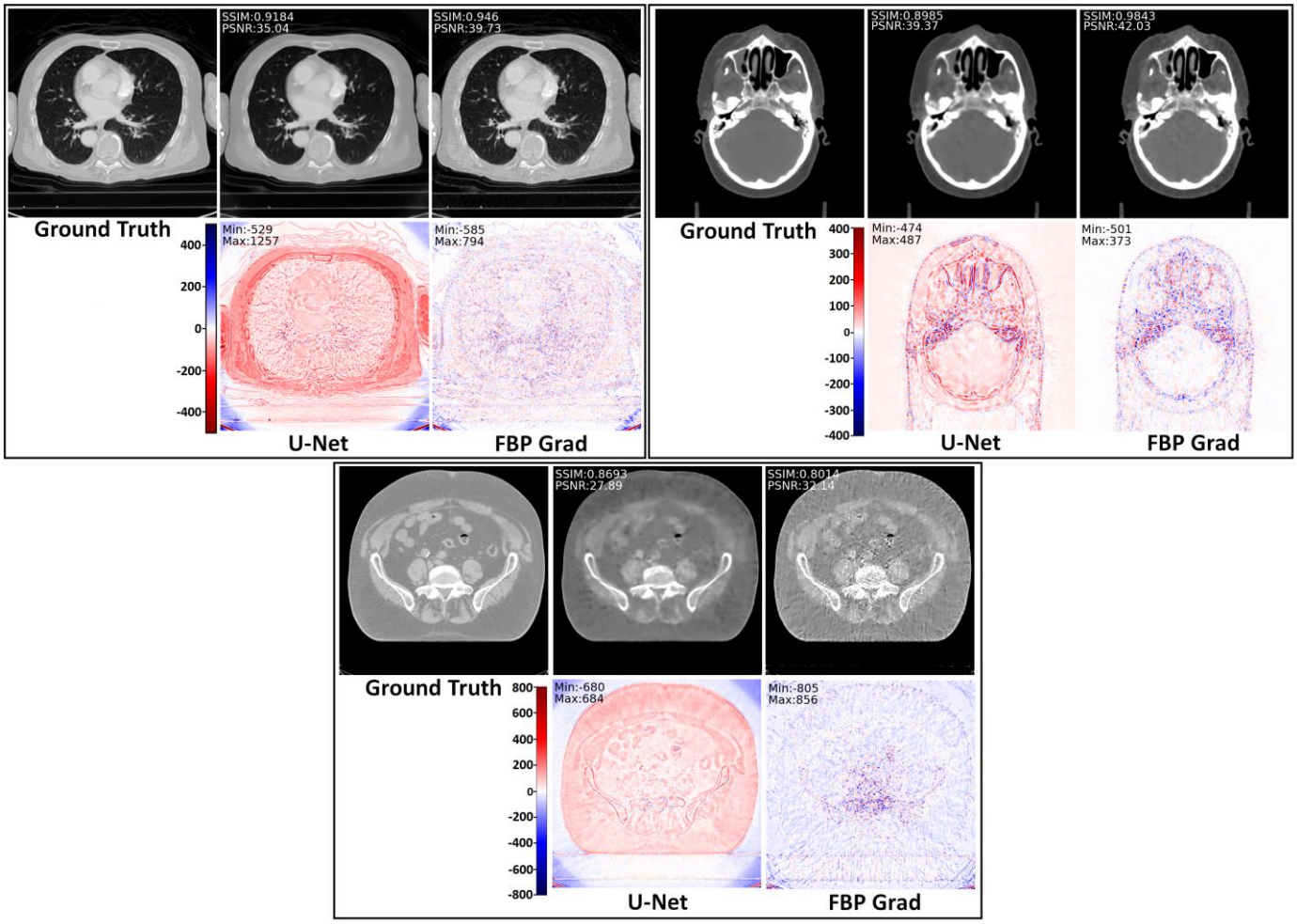

Fig. 14. Sample outputs and the ground truth images along with the error images obtained from the LoDoFaB outputs. The FBP gradient iRIM model performed the best out of the three iRIM models and thus only its outputs are displayed along with the corresponding U-Net outputs for comparison. The same set of ground truth images and image display windows from fig. 12 are used.

Tables containing all the performance evaluation figures calculated on all the results and outputs obtained from each of the iRIM and U-Net models through the course of all the experiments discussed so far can be found in Appendix II.

C. Statistical Test

To further analyse the performance of the iRIM models on each of the above discussed experiments, we use the Wilcoxon signed-rank test provided by python's SciPy library [37]. It is a statistical hypothesis test that can be used for the pairwise comparison of outputs obtained from two models. Here, for each of the above discussed experiments, we compare the SSIM values obtained on the U-Net outputs with the SSIM values obtained on the outputs of the three iRIM models. The null hypothesis is that the median of the paired difference, $SSIM_{U-Net} - SSIM_{iRIM}$ is positive against the alternative that it is negative. P-value close to 0 is interpreted as strong evidence against the null hypothesis. Table 1 shows all the p-values obtained. The p-values that are close to zero and thus that act as strong evidence against the null hypothesis are highlighted using bold text.

It is to be observed that the p-values obtained for all the three iRIM models on the original LoDoPaB test images is zero. This shows that although we did not observe a considerable difference in the mean SSIM and PSNR values earlier, the performance of the iRIM models is statistically significant than the performance of the U-Net on the test data.

The p-values of the adjoint and FBP gradient iRIM models obtained on the head and neck images are all close to zero. On the other hand, the corresponding p-values for pelvic and LoDoPaB test images are 1 for the high noise case and move close to zero for the low noise and LoDoFaB cases. This observation implies that the performances of the iRIM models vary with respect to anatomy and noise contamination level. For the above observation, we neglected the p-values obtained by the adjoint gradient iRIM model on the LoDoFaB dataset. As stated earlier, the reduced performance of the model is suspected to be due to incompatibility issues between the model and fan beam CT operators.

D. Slice-wise Performance Variation

Plots as shown in Fig. 15 are created to visualize the slice-wise performances of the models and to investigate how the performance of the models vary with respect to the body's anatomy with each slice. The first plot belongs to the head and neck scans of one randomly chosen patient from the other anatomy dataset. The SSIM values computed on the corresponding model outputs are arranged in the order of the

TABLE 1. P-values obtained from the Wilcoxon signed-rank test.

Dataset	LoDoPaB Test Images			Head and Neck Images			Pelvic Images		
	Image Grad	Adjoint Grad	FBP Grad	Image Grad	Adjoint Grad	FBP Grad	Image Grad	Adjoint Grad	FBP Grad
Original	**0.0**	**0.0**	**0.0**	1.0	**2.07×10^{-217}**	**1.16×10^{-291}**	1.0	1.0	1.0
High Noise	1.0	1.0	1.0	**5.30×10^{-09}**	**2.98×10^{-93}**	**1.38×10^{-284}**	1.0	1.0	1.0
Low Noise	1.0	**0.0**	**3.54×10^{-186}**	1.0	**0.0**	**0.0**	1.0	**2.46×10^{-187}**	**2.70×10^{-80}**
LoDoFaB	1.0	1.0	**1.90×10^{-22}**	**0.0**	1.0	**0.0**	0.99	1.0	**1.14×10^{-26}**

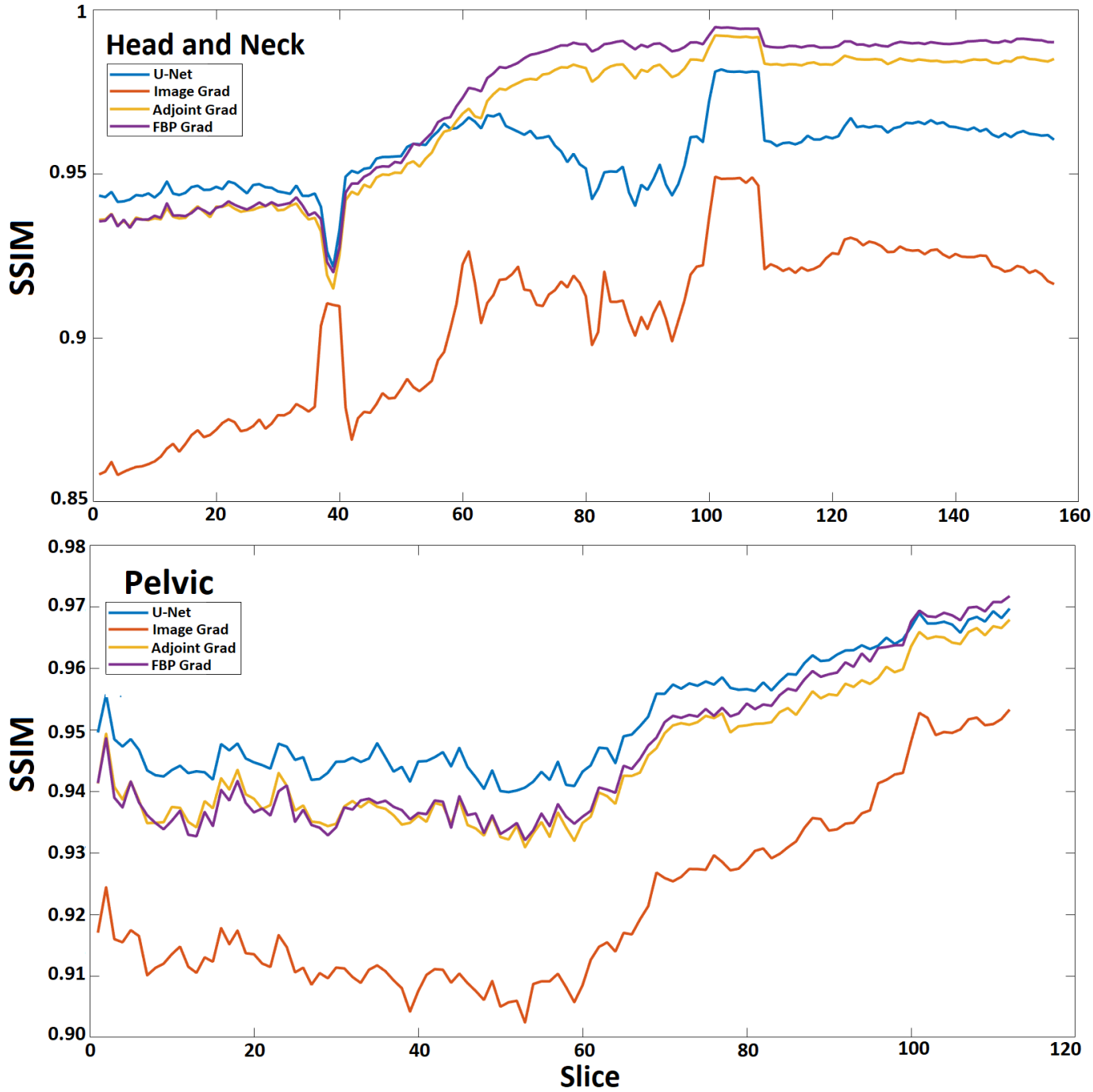

Fig. 15. SSIM values computed on the model outputs ordered slice-wise and plotted against the same. The first plot represents the results obtained on a randomly chosen patient's head and neck scan. The scan begins at the patient's neck region (slice zero) and moves towards the patient's head with each passing slice. The second plot is the results obtained on another randomly chosen patient's pelvic scan. The scan starts at the patient's hip region and moves towards the patient's thighs.

slices and plotted as shown in Fig. 15. Slice zero is a scan at the patient's neck region and the scanner covers a width of 2 mm while moving towards the patient's head with each passing slice. Similarly, the second plot contains the SSIM values computed on the model outputs of a randomly chosen patient's pelvic scans. Slice zero here belongs to a scan close to the patient's hip region and the scanner covers a width of 3 mm while moving towards the patient's thighs with each passing slice.

Both the plots in fig. 15 clearly show that the performances of the adjoint and FBP gradient iRIM models are superior to that of the image gradient iRIM model. Apart from that, a pattern on how the performances of the adjoint and FBP iRIM models vary with respect to the U-Net could be observed. On the first plot in Fig. 15, the U-Net can be seen to have a slight edge on the slices close to the neck region, but as the plot moves towards the head region, both the iRIM models outperform the U-Net by a significant margin. On the second pelvic plot, the U-Net outputs can be seen to have higher SSIM values than the outputs from both the iRIM models in the beginning. However, as we move towards the thighs region, the performances of the iRIM models seem to improve and towards the very end the FBP gradient iRIM model is also able to surpass the U-Net performance by a small margin.

V. Discussion

In this work, three iRIM models for low-dose CT reconstruction were designed with different likelihood gradient complexities and were trained uniformly on the LoDoPaB dataset. The performance of the trained models was initially tested on the LoDoPaB test and challenge dataset. Following that the generalization capabilities of the models to different scenarios were also tested.

Results obtained show that the adjoint and FBP gradient iRIM models outperform the image gradient iRIM models in all the experiments and evaluation tests. Thus, it is evident that likelihood gradient definitions with CT operators have a clear advantage over the likelihood gradients that do not. The CT forward operators within the likelihood gradient feed in relevant information for better reconstructions.

All the three iRIM models were able to outperform the U-Net when tested on the LoDoPaB test data. The p-values obtained from the Wilcoxon test also showed that the performances of the iRIM models on the LoDoPaB test data were statistically significant than that of the U-Net. Performances of our models with respect to other low-dose CT reconstruction approaches are put in perspective by the results that we obtained on the LoDoPaB challenge.

Considering the performance figures from the generalization capability test across anatomies, the iRIM models were able to extend their performance to the head and neck CT data but failed to show performances of adequate quality on the pelvic data. Similar conclusions can be drawn from the results of the high noise dataset too. The iRIM models were able to give good results on the high noise head and neck images while their performances on the high noise Pelvic and LoDoPaB test images were subpar. The performance of the U-Net seems to be reasonably consistent throughout but the iRIM models were able to gain an edge over the performance of the U-Net on the low noise dataset. The iRIM models were able to produce good results on CT data across all three anatomies. The quality of the iRIM pelvic outputs improved significantly and the improvement is also visually evident through the model outputs displayed in fig. 12.

The head is an anatomical region that is dominated by bony structures that absorb much of the x-ray radiation and on the other hand, the pelvic region is predominately composed of soft tissues that allow x-rays to pass through them. Also, there is a significant difference in the amount of tissue the x-rays traverse through in the head and then in the pelvic regions. The larger body contour and the lesser x-ray absorption in the pelvic region might contribute to noisier projection data due to increased photon scattering and other unwanted photon interactions. Since the iRIM model is able to produce good results on head images while being unable to extend it to pelvic, it could be hypothesized that the iRIM model does have the ability to generalize but the prior that has been learnt by the model is only strong enough to handle noise contaminations at low levels. The LoDoPaB test images also have a considerable amount of soft tissue distribution and large body contour. Fig. 11 and the corresponding p-value from table 1 shows that the U-Net was able to perform better than the iRIM models on their high noise simulations. Thus proving the stated hypothesis. Additionally, the improved performance of the iRIM models on the low noise dataset and the pattern that was observed in the plots of Fig. 15 and described in section IV.D also supports this hypothesis.

On the cross-task generalization capability test, the adjoint gradient iRIM model falls apart due to model and fan-beam CT operators incompatibility but the other iRIM models hold their performances. A significant observation was that the performance of the U-Net degraded while the FBP gradient iRIM model held its performance on the LoDoFaB dataset. Additionally, in support of the earlier stated hypothesis that the prior learnt by the iRIM model was only strong enough to handle noise contaminations at low levels. The results obtained on the head and neck images are superior to the ones obtained on the pelvic or thoracic (LoDoPaB) images.

So far all the results were in terms of the two image evaluation metrics – SSIM and PSNR. Through Fig. 16 we try to focus beyond the SSIM and PSNR values and try to make deeper comparisons between the model outputs. A close look at the model outputs on the first row shows that the U-Net and the image gradient iRIM model outputs have moderately blotchy appearances. Unwanted smoothening effects can be noticed, especially on the spine. This effect can be seen to reduce as we move to the outputs of the adjoint and the FBP gradient iRIM models. Additionally, the zoomed versions in the second row of Fig. 16 also show that there is loss in information which decreases with increase in model complexity. The loss is minimum in the FBP gradient iRIM model output. The significance of the likelihood gradient definitions is evident from the observations stated above.

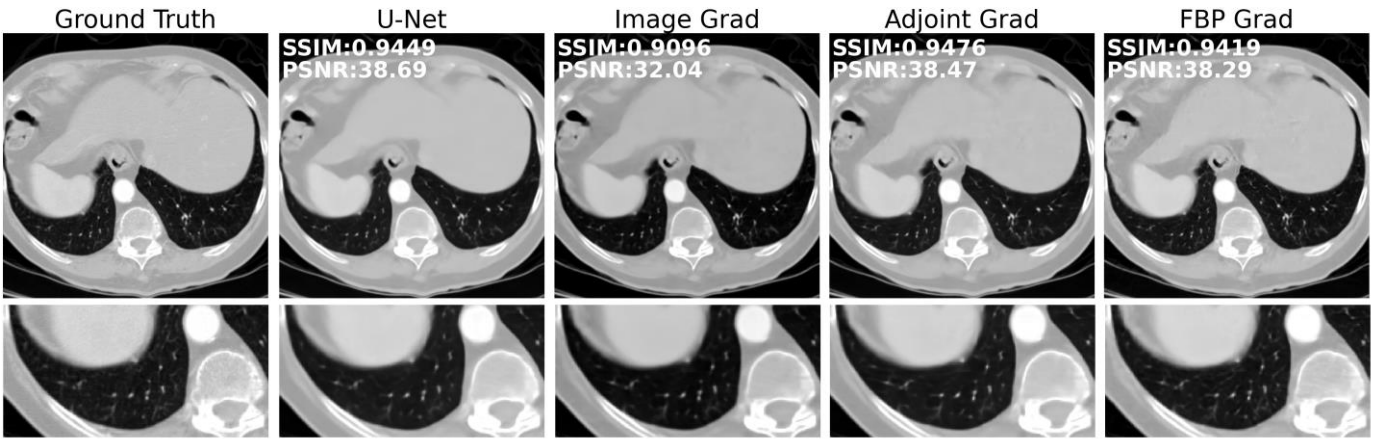

Fig. 16. Another set of outputs along with its ground truth image from the LoDoPaB test data. Slight blotchy or patchy appearance of the U-net and image gradient iRIM model outputs can be noted. On the second row is the zoomed version of the outputs. Unnecessary smoothening effect on the spine along with loss in information can be observed and in the iRIM model outputs the loss and the smoothening effect can be found to decrease with increase in model complexity.

To conclude, the performance of the U-Net was fairly robust and consistent throughout the anatomy and noise level generalization tests. On the iRIM front, the results obtained on the LoDoPaB dataset and the head and neck CT data were notably good. It was also able to give a satisfactory performance on the LoDoFaB dataset showing its cross-task generalization capability. Problems emerged when the model needed to handle higher noise levels as in the case of pelvic images. This could be surpassed by equipping the model to learn an appropriately stronger prior. On that aspect, hyperparameter tuning might be a route to explore. The reference U-Net proved to be a fairly robust model with just 610K learnable parameters. Comparatively, the iRIM model is a huge DL framework with approximately 275M parameters. Although this did not lead to any notable overfitting, there are chances for this to be the reason for the slack in the generalization abilities of the iRIM models. Thus, reducing the iRIM model size might also help improve performance. On that aspect, an attempt to incorporate the U-Net structure into the iRIM framework could also be another step towards performance enhancement. Augmented training data with a wider assortment of CT data can also help the model learn a strong prior. Finally, a recommendation to overcome the limitations of the image quality evaluation metrics would be opting to use more practical tests relating to imaging tasks like segmentation or organ detection. In this way, we can actually quantify the usability of the model outputs instead of entirely depending on SSIM or PSNR values.

VI. Conclusion

Three iRIM models were designed and trained for low-dose CT reconstruction and the generalization capabilities of the trained models across CT data collected over different anatomies, low-dose simulations at different noise levels and different x-ray source beam geometries were tested. From the results, it could be concluded that iRIMs most certainly have the ability to perform low-dose CT reconstructions but there are still a few scopes of improvement that can enhance the overall robustness of the model.

Acknowledgement

This thesis was carried out in UMC, Utrecht. All the pelvic and head & neck CTs that were used during the experiments were from UMC's repository. A huge thanks to my supervisors and my graduation exam committee members for their guidance and valuable support throughout.

References

[1] Kalender, W.A., 2006. X-ray computed tomography. Physics in Medicine & Biology, 51(13), p.R29.
[2] Cassoobhoy, A., 2021. What Is a CT Scan?. [online] WebMD. Available at: <https://www.webmd.com/cancer/what-is-a-ct-scan> [Accessed 27 March 2021].
[3] Eurostat, S.E., Healthcare resource statistics-technical resources and medical technology. 2019. Available from: <https://ec.europa.eu/eurostat/statistics-explained/pdfscache/37388.pdf> [Accessed 27 March 2021].
[4] Gorter, R., Risks of medical radiation. [online] Available at: <http://robert-gorter.info/risks-of-medical-radiation/> [Accessed 27 March 2021].
[5] De González, A.B., Mahesh, M., Kim, K.P., Bhargavan, M., Lewis, R., Mettler, F. and Land, C., 2009. Projected cancer risks from computed tomographic scans performed in the United States in 2007. Archives of internal medicine, 169(22), pp.2071-2077.
[6] Wang, G., Zhang, Y., Ye, X. and Mou, X., 2019. Machine learning for tomographic imaging. IOP Publishing.
[7] Boas, F. E., & Fleischmann, D. (2012). CT artifacts: causes and reduction techniques. Imaging Med, 4(2), 229-240.
[8] Maier, A., Syben, C., Lasser, T. and Riess, C., 2019. A gentle introduction to deep learning in medical image processing. *Zeitschrift für Medizinische Physik*, *29*(2), pp.86-101.
[9] Yedder, H.B., Cardoen, B. and Hamarneh, G., 2020. Deep learning for biomedical image reconstruction: A survey. Artificial Intelligence Review, pp.1-37.
[10] Wang, G., Ye, J.C., Mueller, K. and Fessler, J.A., 2018. Image reconstruction is a new frontier of machine learning. IEEE transactions on medical imaging, 37(6), pp.1289-1296.
[11] Putzky, P. and Welling, M., 2017. Recurrent inference machines for solving inverse problems. arXiv preprint arXiv:1706.04008.
[12] Putzky, P. and Welling, M., 2019. Invert to learn to invert. arXiv preprint arXiv:1911.10914.
[13] Zbontar, J., Knoll, F., Sriram, A., Murrell, T., Huang, Z., Muckley, M.J., Defazio, A., Stern, R., Johnson, P., Bruno, M. and Parente, M., 2018. fastMRI: An open dataset and benchmarks for accelerated MRI. arXiv preprint arXiv:1811.08839.
[14] Putzky, P., Karkalousos, D., Teuwen, J., Miriakov, N., Bakker, B., Caan, M. and Welling, M., 2019. i-RIM applied to the fastMRI challenge. arXiv preprint arXiv:1910.08952.

[15] Leeuwen, T. and Brune, C., 2021. Welcome to Inverse Problems and Imaging — 10 Lectures on Inverse Problems and Imaging. [online] Tristanvanleeuwen.github.io. Available at: <https://tristanvanleeuwen.github.io/IP_and_Im_Lectures/intro.html> [Accessed 27 March 2021].
[16] Gordon, R., Bender, R. and Herman, G.T., 1970. Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography. Journal of theoretical Biology, 29(3), pp.471-481.
[17] Geyer, L.L., Schoepf, U.J., Meinel, F.G., Nance Jr, J.W., Bastarrika, G., Leipsic, J.A., Paul, N.S., Rengo, M., Laghi, A. and De Cecco, C.N., 2015. State of the art: iterative CT reconstruction techniques. Radiology, 276(2), pp.339-357.
[18] Willemink, M.J., de Jong, P.A., Leiner, T., de Heer, L.M., Nievelstein, R.A., Budde, R.P. and Schilham, A.M., 2013. Iterative reconstruction techniques for computed tomography Part 1: technical principles. European radiology, 23(6), pp.1623-1631.
[19] Chen, H., Zhang, Y., Zhang, W., Liao, P., Li, K., Zhou, J. and Wang, G., 2017. Low-dose CT via convolutional neural network. Biomedical optics express, 8(2), pp.679-694.
[20] Chen, H., Zhang, Y., Kalra, M.K., Lin, F., Chen, Y., Liao, P., Zhou, J. and Wang, G., 2017. Low-dose CT with a residual encoder-decoder convolutional neural network. IEEE transactions on medical imaging, 36(12), pp.2524-2535.
[21] Kang, E., Min, J. and Ye, J.C., 2017. A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction. Medical physics, 44(10), pp.e360-e375.
[22] Wolterink, J.M., Leiner, T., Viergever, M.A. and Išgum, I., 2017. Generative adversarial networks for noise reduction in low-dose CT. IEEE transactions on medical imaging, 36(12), pp.2536-2545.
[23] Lossau, T., Nickisch, H., Wissel, T., Morlock, M. and Grass, M., 2020. Learning metal artifact reduction in cardiac CT images with moving pacemakers. Medical image analysis, 61, p.101655.
[24] Claus, B.E., Jin, Y., Gjesteby, L.A., Wang, G. and De Man, B., 2017. Metal-artifact reduction using deep-learning based sinogram completion: initial results. Fully3D 2017 Proceedings, pp.631-635.
[25] Lee, D., Choi, S. and Kim, H.J., 2019. High quality imaging from sparsely sampled computed tomography data with deep learning and wavelet transform in various domains. Medical physics, 46(1), pp.104-115.
[26] He, J. and Ma, J., 2019, March. Radon inversion via deep learning. In Medical Imaging 2019: Physics of Medical Imaging (Vol. 10948, p. 1094810). International Society for Optics and Photonics.
[27] Zhu, B., Liu, J.Z., Cauley, S.F., Rosen, B.R. and Rosen, M.S., 2018. Image reconstruction by domain-transform manifold learning. Nature, 555(7697), pp.487-492.
[28] Moraru. A., 2020. Iterative computed tomography reconstruction using deep learning, Available at: <https://essay.utwente.nl/85407/1/Moraru_MA _EEMCS.pdf >
[29] Adler, J. and Öktem, O., 2018. Learned primal-dual reconstruction. IEEE transactions on medical imaging, 37(6), pp.1322-1332.
[30] Gomez, A.N., Ren, M., Urtasun, R. and Grosse, R.B., 2017. The reversible residual network: Backpropagation without storing activations. arXiv preprint arXiv:1707.04585.
[31] Leuschner, J., Schmidt, M., Baguer, D.O. and Maaß, P., 2019. The lodopab-ct dataset: A benchmark dataset for low-dose ct reconstruction methods. arXiv preprint arXiv:1910.01113.
[32] Adler, J., Kohr, H. and Oktem, O., 2017. Operator discretization library (odl). Software available from https://github. com/odlgroup/odl.
[33] Wirgin, A., 2004. The inverse crime. arXiv preprint math-ph/0401050.
[34] Leuschner, J., Technical reference for the LoDoPaB-CT dataset. Available from: https://github.com/jleuschn/lodopab_tech_ref
[35] Van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E. and Yu, T., 2014. scikit-image: image processing in Python. PeerJ, 2, p.e453.
[36] Leuschner, J., Schmidt, M., Baguer, D.O., Mateus, B. and Erzmann, D., Deep Inversion Validation Library. Available from: https://github.com/jleuschn/dival/tree/master/dival
[37] Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., and SciPy 1.0 Contributors. (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17(3), 261-272. Available from: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wilcoxon.html

Appendix I

Invertible Neural Networks:

A neural layer can be made invertible by splitting its input and output into two parts, $x = (x_1, x_2)$ and $y = (y_1, y_2)$ and by modifying its architecture according to equations (1) to (4) where $\mathcal{F}$ represents a function modelling the layer that is being made invertible. Equations (1) and (2) are the forward computations and equations (3) and (4) are the backward computations.

$$y_1 = x_1 \quad (1)$$

$$y_2 = x_2 + \mathcal{F}(y_1) \quad (2)$$

$$x_2 = y_2 - \mathcal{F}(y_1) \quad (3)$$

$$x_1 = y_1 \quad (4)$$

Invertible Recurrent Inference Machines (iRIM):

These are architecturally modified RIMs that are invertible. Apart from memory saving these frameworks are also equipped to remove training instabilities and have the ability to accommodate large volumes of training data.

The challenge in making RIM invertible is that h_ϕ takes three inputs $(\nabla_{y|x}, x_t, s_t)$ while giving only two outputs (x_t, s_t). The authors overcome this problem by introducing a function g and modifying the update equations as shown below. Equations (5) to (7) are the forward computations and the equations (8) to (10) are the reverse computations. It is to be noted that $\nabla_{y|x}$ in equations (6) and (9) are calculated from x_t'. Also, the learnable parameters of h are no longer shared over iterations t as they were in RIM and this results in increased model expressiveness. Fig. 1 shows the forward and the reverse computations involved in an iRIM step.

$$x_t' = x_t \quad (5)$$

$$s_t' = s_t + g\left(\nabla_{y|x}\right) \quad (6)$$

$$x_{t+1}, s_{t+1} = h_t(x_t', s_t') \quad (7)$$

$$x_t', s_t' = h_t^{-1}(x_{t+1}, s_{t+1}) \quad (8)$$

$$s_t = s_t' - g\left(\nabla_{y|x}\right) \quad (9)$$

$$x_t = x_t' \quad (10)$$

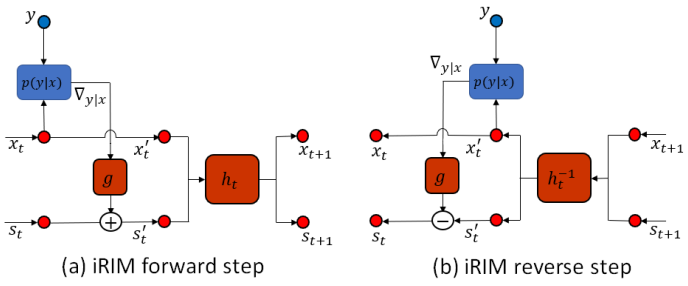

Fig. 1. Forward and reverse step of an iRIM framework. Its correspondence to the update equation (16) to (21) can be noted. Image sourced from [12]

For h_t to be invertible, each layer within it must be made invertible. Putzky et al also introduce this invertible layer and it uses the following update equations.

$$x' = \mathrm{U}x \quad (11)$$

$$y_1' = x_1' \quad (12)$$

$$y_2' = x_2' + \mathcal{G}(x_1') \quad (13)$$

$$y = \mathrm{U}^T y' \quad (14)$$

$$y' = \mathrm{U}y \quad (15)$$

$$x_1' = y_1' \quad (16)$$

$$x_2' = y_2' - \mathcal{G}(y_1') \quad (17)$$

$$x = \mathrm{U}^T x' \quad (18)$$

Equations (11) to (14) are the forward computations and equations (15) to (18) are reverse computations as shown in fig. 2 (a). U is an orthogonal 1×1 convolution and due to its orthogonality its inverse will be equal to its transpose U^T. This $\mathrm{U}^{-1} = \mathrm{U}^T$ property helps reduce computational cost during training. $\mathcal{G}$ is a residual block with three convolution layers as shown in Fig. 2 (b). It takes as input k, the number of channels in the hidden layers and d, the downsampling factor. The first layer performs a spatial downsampling operation along with pixel shuffling using convolution with stride d and filter of size $d \times d$. The second layer performs a 3×3 convolution with stride 1. The last layer is a transpose convolution layer that reverses the downsampling operation of the first layer. There is a ReLU layer added after the first and second convolution layers and a Gated Linear Unit (GLU) at the output.

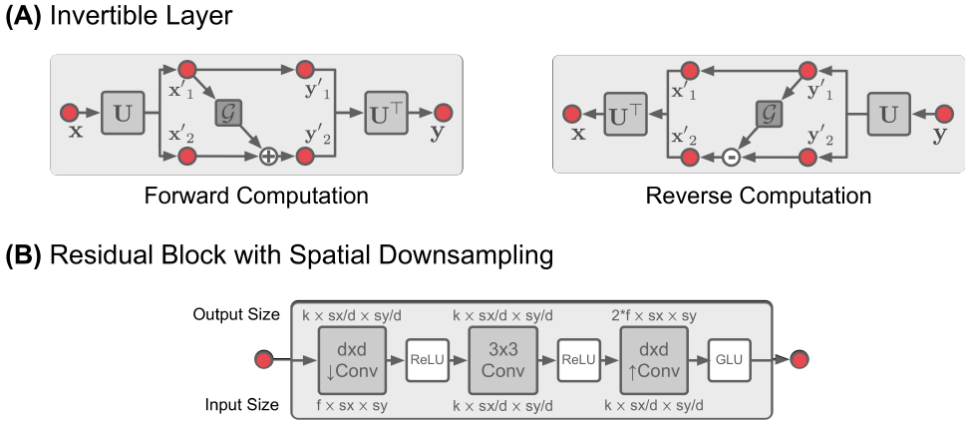

Fig. 2. (a) The invertible layer used within an iRIM model. This is a graphical illustration of the equations (2) to (29) and the resemblance can be clearly observed. (b) The function $\mathcal{G}$ defined to be used in the invertible layer. Image sourced from [12]

Appendix II

TABLE 1. LoDoPaB Test Results

	SSIM					PSNR				
	Mean	**Median**	**Std**	**Minimum**	**Maximum**	**Mean**	**Median**	**Std**	**Minimum**	**Maximum**
LD FBP	0.3586	0.3842	0.1466	0.0503	0.647	21.86	22.42	3.67	12.84	30.49
U-Net	0.8443	0.9074	0.1501	0.1818	0.99	35.84	36.65	4.59	20.04	47.97
Image Grad	0.8497	0.9105	0.1458	0.1952	0.9909	35.84	36.68	4.63	20.03	47.89
Adjoint Grad	0.8541	0.912	0.1394	0.2286	0.9918	35.94	36.79	4.74	20.14	48.32
FBP Grad	0.8532	0.9107	0.1391	0.2318	0.9918	35.87	36.74	4.8	20.14	48.32

TABLE 2. LoDoPaB Challenge Results

	SSIM					PSNR				
	Mean	**Median**	**Std**	**Minimum**	**Maximum**	**Mean**	**Median**	**Std**	**Minimum**	**Maximum**
Model 1	0.8683	0.9152	0.1133	0.3597	0.9842	36.36	37.15	3.74	24.59	45.62
Model 2	0.8667	0.9149	0.116	0.3347	0.9863	36.32	37.1	3.75	24.54	45.66
Model 3	0.8659	0.9156	0.1189	0.327	0.9861	36.33	37.14	3.76	24.42	45.58
Adjoint Grad	0.8692	0.9159	0.1122	0.3637	0.9862	36.08	36.89	3.76	22.97	45.26
Image Grad	0.8664	0.9146	0.1155	0.3348	0.9857	35.99	36.77	3.67	24.17	44.95
FBP Grad	0.8602	0.9053	0.1129	0.3441	0.9828	35.48	35.95	3.5	24.35	44.65

TABLE 3. Results obtained on the Other Anatomy dataset (SSIM)

	Head and Neck Images					Pelvic Images				
	Mean	**Median**	**Std**	**Minimum**	**Maximum**	**Mean**	**Median**	**Std**	**Minimum**	**Maximum**
LD FBP	0.6691	0.6992	0.1339	0.3289	0.9159	0.3098	0.2903	0.0979	0.1646	0.6366
U-Net	0.9566	0.9592	0.0175	0.8261	0.9854	0.9507	0.9506	0.0144	0.9169	0.9839
Image Grad	0.9369	0.9419	0.0264	0.8096	0.9816	0.9032	0.9088	0.0332	0.8178	0.9661
Adjoint Grad	0.9718	0.9794	0.0163	0.915	0.9927	0.9423	0.9453	0.0208	0.8832	0.9817
FBP Grad	0.9775	0.9864	0.0176	0.916	0.9965	0.9366	0.9412	0.0264	0.8612	0.9829

TABLE 4. Results obtained on the Other Anatomy dataset (PSNR)

	Head and Neck Images					Pelvic Images				
	Mean	**Median**	**Std**	**Minimum**	**Maximum**	**Mean**	**Median**	**Std**	**Minimum**	**Maximum**
LD FBP	32.28	32.93	3.51	23.41	40.91	23.94	23.76	2.58	18.9	31.87
U-Net	41.83	42.10	2.40	36.18	48.59	40.05	39.66	2.05	35.87	46.00
Image Grad	38.67	39.82	4.00	29.07	48.30	32.32	31.94	2.74	27.2	40.65
Adjoint Grad	42.05	42.44	2.79	35.76	52.49	39.06	38.73	2.35	33.53	45.43
FBP Grad	42.56	42.99	2.94	35.85	53.96	39.01	38.71	2.57	33.32	45.74

TABLE 5. Results obtained on the High Noise dataset (SSIM)

	LoDoPaB Test Images					Head and Neck Images					Pelvic Images				
	Mean	**Median**	**Std**	**Min**	**Max**	**Mean**	**Median**	**Std**	**Min**	**Max**	**Mean**	**Median**	**Std**	**Min**	**Max**
LD FBP	0.2886	0.3056	0.1446	0.0226	0.6224	0.5665	0.5932	0.1397	0.2338	0.8476	0.2289	0.2025	0.0827	0.1199	0.546
U-Net	0.8079	0.8612	0.15	0.1845	0.9797	0.924	0.9239	0.0244	0.7814	0.9724	0.9317	0.9319	0.0148	0.8908	0.9756
Image Grad	0.6689	0.7072	0.1532	0.2144	0.9307	0.9273	0.9347	0.0336	0.8042	0.9796	0.8712	0.8742	0.0441	0.7411	0.9592
Adjoint Grad	0.6776	0.736	0.2025	0.1819	0.9515	0.9453	0.9644	0.0374	0.7832	0.9878	0.8332	0.8474	0.0821	0.5686	0.9607
FBP Grad	0.7684	0.8089	0.1476	0.2106	0.9609	0.967	0.9804	0.0261	0.8721	0.9958	0.9042	0.9098	0.0414	0.7771	0.9743

TABLE 6. Results obtained on the High Noise dataset (PSNR)

	LoDoPaB Test Images					Head and Neck Images					Pelvic Images				
	Mean	**Median**	**Std**	**Min**	**Max**	**Mean**	**Median**	**Std**	**Min**	**Max**	**Mean**	**Median**	**Std**	**Min**	**Max**
LD FBP	20.53	21.39	4.50	9.71	30.28	29.87	30.61	3.63	20.86	38.14	21.32	20.92	2.55	16.27	29.78
U-Net	33.93	34.41	4.16	19.92	45.37	40.51	40.82	2.44	34.96	47.27	38.54	38.18	2.03	33.80	44.67
Image Grad	26.62	26.85	3.79	18.12	36.88	37.3	38.51	4.19	26.75	47.84	30.34	29.85	2.77	25.27	39.18
Adjoint Grad	29.29	30.22	5.24	17.63	40.03	39.72	40.30	3.39	31.02	51.01	34.04	33.86	3.50	25.02	42.14
FBP Grad	32.15	32.31	3.87	19.69	41.44	40.86	41.34	3.17	33.56	52.85	36.84	36.49	2.77	30.52	43.91

TABLE 7. Results obtained on the Low Noise dataset (SSIM)

	LoDoPaB Test Images					Head and Neck Images					Pelvic Images				
	Mean	**Median**	**Std**	**Min**	**Max**	**Mean**	**Median**	**Std**	**Min**	**Max**	**Mean**	**Median**	**Std**	**Min**	**Max**
LD FBP	0.5192	0.5636	0.1785	0.102	0.8411	0.7804	0.8126	0.1097	0.4614	0.9549	0.4529	0.4271	0.1029	0.2972	0.777
U-Net	0.8512	0.9163	0.1519	0.1716	0.9912	0.9689	0.9743	0.0173	0.8472	0.991	0.9553	0.9551	0.0135	0.9238	0.9865
Image Grad	0.8155	0.867	0.1469	0.1865	0.9755	0.9458	0.949	0.0204	0.8179	0.9824	0.9302	0.9356	0.0225	0.8755	0.9754
Adjoint Grad	0.8559	0.9157	0.1447	0.2301	0.9923	0.9783	0.9835	0.0119	0.9377	0.9942	0.9593	0.9606	0.0124	0.9262	0.9866
FBP Grad	0.856	0.913	0.1446	0.2012	0.9911	0.9832	0.9898	0.0128	0.9393	0.9967	0.9576	0.959	0.0153	0.9182	0.9875

TABLE 8. Results obtained on the Low Noise dataset (PSNR)

	LoDoPaB Test Images					Head and Neck Images					Pelvic Images				
	Mean	**Median**	**Std**	**Min**	**Max**	**Mean**	**Median**	**Std**	**Min**	**Max**	**Mean**	**Median**	**Std**	**Min**	**Max**
LD FBP	26.77	27.43	3.95	17.71	35.81	35.24	35.78	3.32	27.16	43.55	27.49	27.07	2.43	23.16	35.31
U-Net	36.51	37.45	4.91	20.00	49.21	42.75	42.97	2.34	37.31	49.48	40.99	40.63	2.06	36.99	47.14
Image Grad	31.29	31.45	3.85	19.31	41.02	40.00	40.97	3.60	31.19	48.5	35.11	34.64	2.51	30.23	42.50
Adjoint Grad	36.23	37.05	4.76	20.18	48.58	43.21	43.5	2.55	37.43	52.89	40.92	40.59	1.99	36.8	47.01
FBP Grad	36.40	37.15	4.84	20.08	48.50	43.81	44.21	2.72	37.77	54.39	41.03	40.65	2.25	36.53	47.38

TABLE 9. Results obtained on the LoDoFaB dataset (SSIM)

	LoDoPaB Test Images					Head and Neck Images					Pelvic Images				
	Mean	**Median**	**Std**	**Min**	**Max**	**Mean**	**Median**	**Std**	**Min**	**Max**	**Mean**	**Median**	**Std**	**Min**	**Max**
LD FBP	0.3676	0.3931	0.1606	0.0441	0.7219	0.3676	0.6662	0.7096	0.1402	0.3153	0.9013	0.3097	0.289	0.0876	0.1751
U-Net	0.7266	0.7792	0.149	0.1402	0.9194	0.7266	0.7903	0.8109	0.096	0.4911	0.9476	0.8461	0.8636	0.0531	0.6617
Image Grad	0.6069	0.6374	0.1508	0.2005	0.9204	0.6069	0.9166	0.9232	0.0391	0.7887	0.9775	0.8419	0.8428	0.0439	0.7125
Adjoint Grad	0.0511	0.0468	0.0292	0.01	0.1611	0.0511	0.0585	0.0549	0.0212	0.0231	0.143	0.0273	0.0242	0.0087	0.0174
FBP Grad	0.7268	0.7691	0.1602	0.2101	0.9558	0.7268	0.9644	0.9809	0.0299	0.8518	0.9933	0.8823	0.8908	0.0544	0.7298

TABLE 10. Results obtained on the LoDoPaB dataset (PSNR)

	LoDoPaB Test Images					Head and Neck Images					Pelvic Images				
	Mean	**Median**	**Std**	**Min**	**Max**	**Mean**	**Median**	**Std**	**Min**	**Max**	**Mean**	**Median**	**Std**	**Min**	**Max**
LD FBP	22.21	22.66	3.68	13.17	31.66	31.73	32.74	3.75	22.71	39.54	23.97	23.60	2.43	19.48	31.07
U-Net	26.6	26.17	2.99	18.82	37.65	36.48	38.04	4.16	27.24	45.18	30.74	30.69	2.29	25.33	37.54
Image Grad	23.00	23.04	3.44	15.31	34.68	35.23	37.33	5.08	23.2	45.68	26.89	26.52	2.87	20.98	34.34
Adjoint Grad	13.12	12.90	2.10	8.13	19.00	17.13	17.28	2.66	11.37	28.32	12.55	12.27	1.30	10.26	17.93
FBP Grad	30.14	30.15	4.13	19.34	41.14	41.33	42.10	3.59	31.97	51.71	35.90	35.74	2.76	29.54	43.02