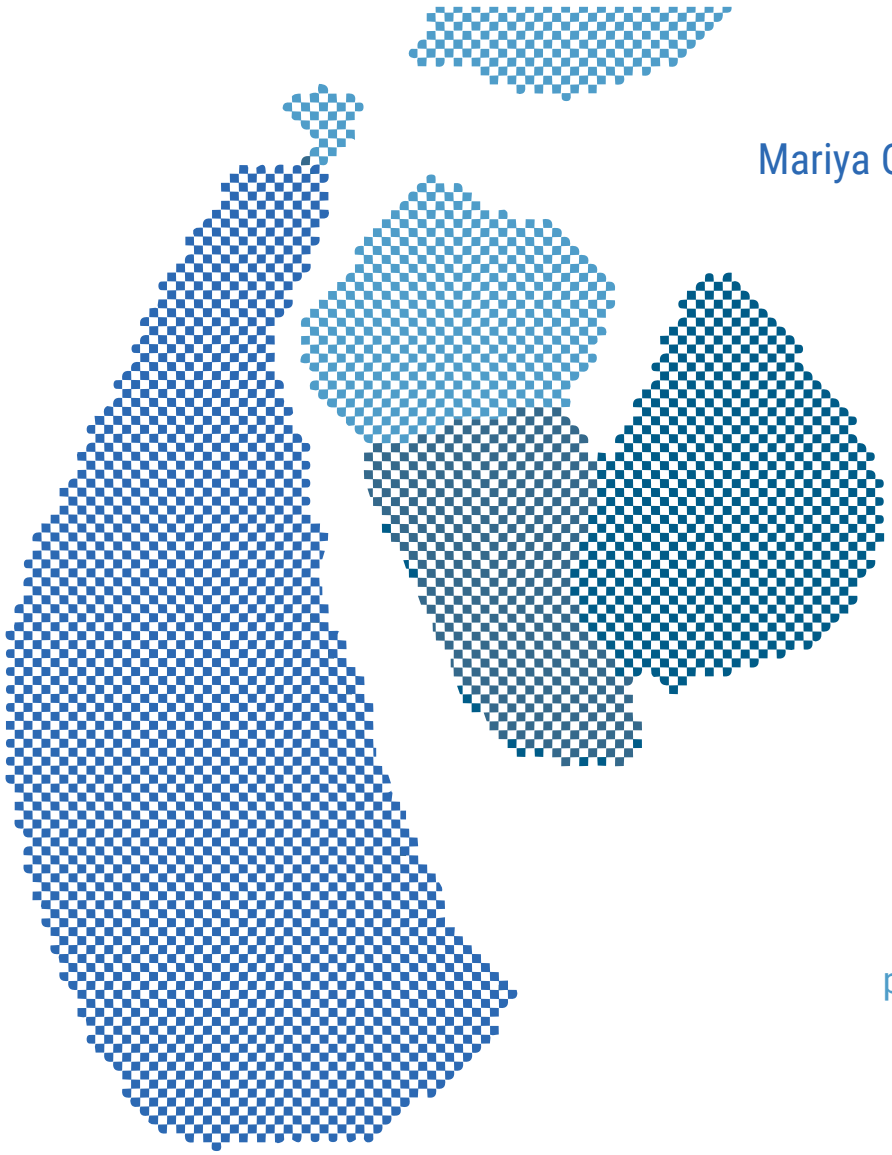
A Data-Driven Model for Direct Segmentation in Computed Tomography Imaging



Mariya Olegivna Karlashchuk

MSc Thesis

September 2022

Assessment committee

prof. dr. Christoph Brune

dr. Jelmer Wolterink

dr. Katharina Proksch

Mathematics of Imaging & AI

Faculty of Electrical Engineering, Mathematics and Computer Science

UNIVERSITY OF TWENTE.

Abstract

Computed tomography (CT) is a powerful tool in medical imaging that is used to study anatomical structures in the human body. To achieve this, CT measurement data is reconstructed as an image, after which semantic segmentation is performed to analyse body structures or to detect diseases. These so-called sequential methods perform reconstruction and segmentation separately and are frequently used in practice, just as joint methods which perform these tasks simultaneously. However, these methods have several disadvantages. The reconstructed images could be not suited for segmentation or might contain noise that results in incorrectly predicted segmentation masks, which might be fatal in practical applications. Next to this, reconstructing an image is not necessary to detect the location of a certain body substructure, since measurement data contains this information too. In this thesis, a direct deep learning model is evaluated that performs segmentation using extracted features from measurement data. The model learns information based on the geometry of the measurements and uses this to predict the relation between sinusoids in sinograms and coordinates in a corresponding segmentation mask. The direct segmentation model is compared to a joint model, which is adapted from the aforementioned geometry-based model. The joint model reconstructs images and predicst segmentation masks based on the same learned relation. Both models are assessed on their performance in segmentation and reconstruction.

***Keywords**: CT, image reconstruction, semantic segmentation, multi-task learning, direct method*

Preface

Before you lies my thesis titled 'A Data-Driven Method for Direct Segmentation in Computed Tomography imaging'. In the past year I have put all my focus and dedication in the evaluation of a novel method that has potential for ground-breaking research in the future concerning medical imaging of CT. Applying my knowledge of deep learning in the medical field has broadened my view of where mathematics can be applied in practice and how to adapt to this new environment. Next to this, it was exciting, yet sometimes cumbersome to develop myself as a researcher who is (hopefully) worthy of a Master of Science degree. I am very happy that I have completed this journey in my studies, which has taught me more things than I could even think of. If you are close to me, you might know that the past year has not been kind to me. As unknowing reader you must know that the remainders of a certain pandemic combined with certain personal circumstances hindered my progress in research. During these times I am glad I had people around me who could either support me, give advice or just understand the situation. Finish my studies as a nine-years student instead of an eight-years was something that I did not expect, but nevertheless I am glad to end the Master Applied Mathematics with this masterpiece (no pun intended).

Just as in every preface, I also must, want and will to thank many people. The first of them being my daily supervisor, Jelmer Wolterink. Thank you very much for your supervision, support, and understanding. I am very grateful for all your help, especially the late-night questions which were answered with a late-night response. When I started my thesis, I enjoyed your enthusiasm and how it sparkled my own enthusiasm about the subject. I remember the moments when I did not feel as enthusiastic about my research and how everything was going and I still am surprised at how light and happy I felt after a meeting with you. You have taught me to trust my own instinct in research, to not feel insecure and be satisfied with all my achievements. That is something I am very thankful for and something that I will probably use quite often in the future.

Secondly, I would like to thank my second and third supervisors, Christoph Brune and Katharina Proksch. Thank you very much for your supervision and your critical views and feedback that you provided me of during our green-light meeting. Your questions made me do something that I am insecure about and had to learn to do propely: relating things to mathematics. Data Science was my safe-space, where you could easily brush over the math and focus more on machine learning itself. However, thanks to you I have dared myself to relate my findings and put practical applications to a mathematical context.

Next, I want to thank Dieuwertje and Julian for being my unofficial sort-of fourth and fifth supervisors. I am glad that the door was always open for me and I could ask questions at any moment. Without your help, I would probably know nothing about constructing a dataset in MONAI and not be able to train my network on the GPU's. Besides, the informal drinks and chats were also great to experience with you.

Thank you, Chris, as well for your immense support. I have no idea what I would do without your help. You were always there for me, during all kinds of moments during my research. Mom, dad, Vasiliy: also a huge thanks to you for supporting me, being patient with me and encouraging me during not only this year, but during every academic year. Now I have finally time to call or visit you more.

Lastly, I want to thank all my friends from Euphemia, my mathy friends, and 't Hok-colleagues. You made my final year as math student amazing, I will forever cherish my time here with you guys. The final shout-out goes to Thomas, Justus and Martijn, who helped me proofread my thesis and write it as it is now.

Contents

Chapter 1

Introduction

Computed tomography (CT) is one of the medical imaging techniques in radiology that is widely used to visualize internal structures of the human body [1]. When a CT scan is performed, X-rays are emitted by rotating X-ray tubes and the resulting X-ray attenuation is measured by a detector over different angles. A mathematical reconstruction algorithm provides a cross-sectional image of the body of a patient, where structures such as bones, organs and soft-tissue are visualised with different intensities. By stacking multiple cross-sectional images (slices), a complete CT scan is acquired that provides a three-dimensional image. The spatial resolution of this image is high compared to modalities such as position emission tomography (PET) and magnetic resonance imaging (MRI), which makes CT a very powerful tool in diagnosis, prognosis, and treatment planning [2]. However, performing a CT scan also has its downsides. A CT scan requires ionizing radiation, which can damage body cells of a patient which in its turn might lead to cancer [3]. An obvious yet crucial solution to this issue is to reduce the amount of radiation the patient is exposed to. However, while risks associated to CT decrease, a CT scan with lower radiation dose results in noisy projection data [4]. Mapping this measurement data onto the image domain might lead to more noise appearing in the reconstructed image and thus incorrect visualisation of anatomy and pathology, complicating diagnosis or treatment of a patient.

To work around the issue of noisy projection data, one could incorporate prior knowledge of the target image in reconstruction algorithms. Traditional analytical methods, such as filtered back-projection (FBP) and iterative reconstruction (IR) algorithms, can only handle limited prior information regarding, for example, the anatomy of the human body [5]. On the other hand, machine learning approaches - specifically, deep learning methods - have proven to be very useful techniques for CT image reconstruction tasks [6]. Deep learning methods are able to learn features of the provided data directly, where conventional methods rely on hand-crafted features of the input. Deep learning algorithms are able to create mappings from raw input to a desired output [7], meaning they are well-suited for the reconstruction of noisy medical images since they are able to capture complex non-linear transformations [8].

High-quality image reconstruction is important for a range of applications, among them the automatic segmentation of anatomy and pathology. Image reconstruction and segmentation are two separate research areas that are typically used sequentially in medical imaging and image analysis. Image segmentation is a technique that partitions an image in multiple regions, the so-called segments. The goal is to break down an image into multiple subgroups and thus reduce the complexity of the image to ease any further processing or the analysis of the image. This technique is used in medical imaging, for example to locate anomalies or to study the anatomical structure of part of the patient's body. Deep learning methods are frequently used for segmentation tasks as well [9].

Most methods that are currently developed in CT image analysis can be considered as indirect methods, meaning that first CT projection data (so-called sinogram) is reconstructed resulting in a CT image, after which segmentation is performed on reconstructions. However, most research - especially in deep learning - focuses on only one part of this sequential approach: either the reconstruction model is improved, or the segmentation method. Even though indirect methods are regularly used in practice, they suffer from several disadvantages. A reconstructed image might not necessarily be suited for segmentation, since the reconstruction algorithm might use only a part of the representation of the acquired data [10]. Especially when the radiation dose of a CT scan is reduced, anomalies might be detected incorrectly due to noisy reconstructions [11]. In addition to this, when segmenting part of the

visualised anatomy or pathology, one is interested in a specific region, not the entire reconstruction. Essentially, an adequate reconstruction is not essential to perform correct segmentation. Therefore, an alternative for these sequential approaches is to perform joint end-to-end reconstruction and segmentation of images, where the model trains by means of a combined reconstruction and segmentation loss. This type of method would lead to preservation of information on the reconstructed image when performing segmentation on the image and propagation of the loss through the entire network.

Considering the idea of performing reconstruction and segmentation jointly, the question arises whether it is necessary to even reconstruct an image at all, instead of performing segmentation on the projection data in the measurement domain. The purpose of reconstructed images is, as mentioned before, primarily to support the observations that are passed on from a radiologist to a clinician. Chung et al. [12] discuss this very purpose of reconstructing an image. In practice, one might only be interested in the growth or reduction of a certain disease during a screening, for example. These types of minor procedures do not require a full CT scan including a reconstruction. Therefore, it is worthwhile considering a data-driven method by performing segmentation directly on the sinograms, since the projection data contains the same (and perhaps even more) information as the reconstructed images. The work of Chung et al. mentions that deep learning techniques could contribute to more robust, objective and quantitative methods of the diagnosis and treatment of patients. Features can be derived directly from sinograms, possibly improving the performance of a deep learning model that is suited for segmentation [10].

This thesis explores the potential of a data-driven deep learning model and whether it could overcome the limitations of indirect methods that are currently used in medical imaging. The goal is to determine if such a model is suited for *direct segmentation* on CT measurement data. The proposed model is based on the findings in the work of He et al. [13][14]. There, a downsampled-imaging-geometry based network (DSigNet) for CT image reconstruction is described, which combines geometric modelling knowledge and prior knowledge obtained from a data-driven training process [14]. The DSigNet model is able to exploit geometric knowledge and prior knowledge from sinograms and back-project these from the projection domain onto the image domain. In this work, the DSigNet is adjusted to a direct-DSigNet model to perform direct segmentation and a joint DSigNet model, which performs reconstruction and segmentation simultaneously. The three models are compared to each other to determine if direct segmentation is indeed superior to indirect segmentation and a joint model.

1.1 Thesis outline

Chapter 2 presents the mathematics behind CT acquisition and reconstruction and reviews methods that are currently used in the field of CT. Chapter 3 describes the theory and practices of data-driven machine learning approaches for CT reconstruction. The mathematical models form the basis for the models to be assessed in this research. Chapter 4 presents the proposed model that provides a segmentation using only sinograms as input. A similar model that is able to perform reconstruction and segmentation simultaneously is specified as well. Moreover, the dataset used for experiments is described and the evaluation metrics are provided. Chapter 5 presents the results that were acquired while comparing both methods described in Chapter 4. This is followed by a discussion and points for further research in Chapter 6, and a general conclusion in Chapter 7.

Chapter 2

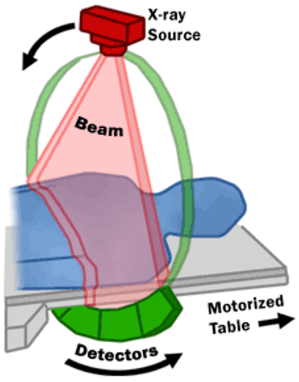
Image Analysis in Computed Tomography

This chapter is an introduction to medical image analysis in CT. First, the physical functioning and the mathematical background of CT scan reconstruction are described in Chapter 2.1 and 2.2, respectively. This is followed by Chapter 2.3 which addresses ill-posedness in image reconstruction in CT, which is a critical issue in image reconstruction and semantic segmentation. Chapter 2.4 discusses conventional and deep learning models that are used in practice which are less prone to ill-posedness in image reconstruction. This is followed by conventional and deep learning models that are used for semantic segmentation in Chapter 2.5. Lastly, Chapter 2.6 elaborates on how image reconstruction models and semantic segmentation models are used sequentially or jointly in practice.

2.1 CT scan acquisition

Computed tomography is a medical imaging technique that is used in medical diagnosis and is performed by a technologist. Before a CT scan is made, a patient is positioned on a motorized table that moves through the scanner, as depicted in Figure 2.1a. During a CT scan, X-ray sources with detectors on the opposing side are both rotating around the patient. The sources emit X-ray beams, which are lined up with detectors after passing through the patient. The beams are emitted in either a fan-beam shape (see Figure 2.1b) or a parallel-beam shape (see Figure 2.3a). In this work, it is assumed the beams are emitted parallel to each other.

(a)



(b)

Figure 2.1: (a) CT scan performed at Leids Universitair Medisch Centrum [15], (b) Schematic image of a fan-beam CT scan performed on a patient [16].

After one full rotation of the source-detector pairs, a sinogram is obtained. One full rotation denotes a rotation over either a semi-circle or a full circle, depending on the type of CT scan. This sinogram can be used to reconstruct a *slice*, which is an image of a cross-section of the scanned body of the patient. The slices depict body tissues and their densities, which are proportional to the X-ray attenuation. This results in low attenuation being displayed as

a low intensity (dark colors) and high attenuation as high intensity (bright colors). For example, bone tissue has a very high attenuation, whereas bodily fluids and blood have a low attenuation. Figure 2.2 shows a sinogram-reconstruction pair where the different attenuations are clearly visible in a slice.

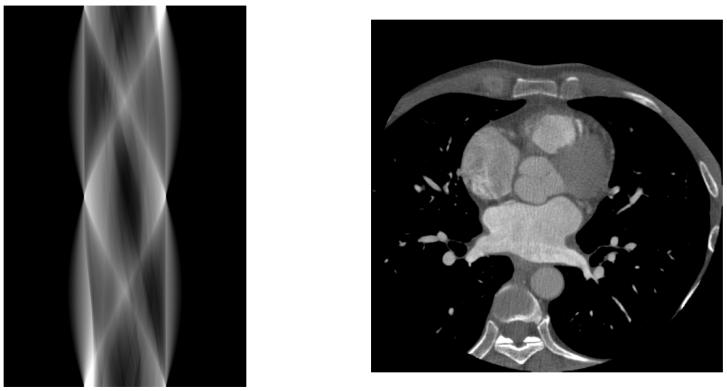


Figure 2.2: Example of a sinogram and a corresponding reconstructed image.

After the first rotation, the table is moved slightly forward, followed by another rotation of the source-detector-pairs. This process is repeated until the desired number of slices is obtained. These slices can be stacked such that a three-dimensional image is obtained of (a part of) the patients body. These images provide a visualization of anatomy and pathology, which can be used for prognosis, diagnosis, and treatment planning, among other applications.

2.2 Mathematics of CT

The following theory is based on the lecture notes provided by Van Leeuwen and Brune [17]. Note that the mathematical derivation for CT is explained for a two-dimensional case. Assume X-rays are emitted from a starting point $x = 0$ with intensity I_0 and detected at distance $x = +\infty$. By the Beer-Lambert law the attenuation in a small interval of length δx corresponds to the difference between the intensity $I(x)$ and the attenuation at x, resulting in the following expression:

$$I(x + \delta x) = I(x) - u(x)I(x)\delta x \tag{2.1}$$

with $u(x)$ being the attenuation coefficient at position x. Eq. 2.1 can be rewritten in the form of a differential equation

$$\frac{I(x + \delta x) - I(x)}{\delta x} = -u(x)I(x), \tag{2.2}$$

with the general solution

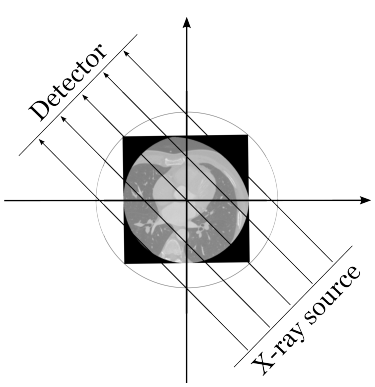
$$I(x) = I_0 \exp\left(-\int_0^x u(t)\mathrm{d}t\right). \tag{2.3}$$

Now, assume that the beams travel along an arbitrary finite ray ℓ, starting at the source and ending at the detector. By defining the measured intensity to be I_m, rewriting Eq. 2.3 gives:

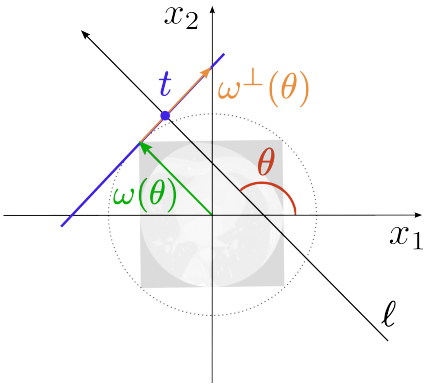
$$P_\ell = -\log\left(\frac{I_m}{I_0}\right) = \int_\ell u(x)\mathrm{d}x, \tag{2.4}$$

which relates the measured and inital intensity and the attenuation coefficient. Note that P_ℓ is known with a certain error, whereas $u(x)$ is unknown. This expression holds for a single X-ray beam.

The idea of CT is to obtain information of the interior of a three-dimensional object by rotating the source and detector around the object in order to obtain and combine multiple projections. This implies that a general expression ought to be derived for the measured intensity for all measurements. By finding a parametrisation for all X-ray lines, a transform is constructed which is expressed as an integral.



(a) Parallel-beam CT scanning. X-rays are emitted from the X-ray source and travel in parallel straight lines through an object until observed by the detector.



(b) Geometrical interpretation of parallel beam CT scanning. The black arrow crossing the object and the unit sphere is the X-ray beam to be parametrised.

Figure 2.3: Schematic and geometric visualisation of a parallel-beam CT scan retrieved by a parallel-beam CT scan.

First, consider the detector and X-ray source to be rotating around a unit circle. A mathematical description of CT will be derived in two spatial dimensions $x = (x_1, x_2) \in \mathbb{R}^2$. Now, assume that the attenuation $u(x_1, x_2)$ of an object at point (x_1, x_2) is constrained to the unit sphere $\{x_1^2 + x_2^2 \leq 1\}$ and the detector is positioned along the tangent of the unit circle (see Figure 2.3b). The angular position of the detector with respect to the x_1-axis is constrained as $\theta \in [0, \pi)$. The point of contact between the detector and the X-ray can be expressed as the normal of this point on the unit sphere $\omega(\theta)$, while the normal of $\omega(\theta)$, i.e. $\omega^{\perp}(\theta)$, is the unit vector along the tangent. Any point on the detector line can now be described as a linear function $\omega(\theta) + t\omega^{\perp}(\theta)$ with parametrisation $t \in \mathbb{R}$. Translating this to an X-ray line ℓ that is orthogonal to the detector, which it crosses at a point corresponding to $t \in [-1, 1]$, this line can now be parametrised by $s \in \mathbb{R}$ as follows:

$$x(s) = s\omega(\theta) + t\omega^{\perp}(\theta) = s \begin{bmatrix} -\sin(\theta) \\ \cos(\theta) \end{bmatrix} + t \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} \tag{2.5}$$

By rewriting Eq. 2.5, an expression can be derived for all points lying on ℓ:

$$\ell(\theta, t) = \{x \in \mathbb{R} \mid x_1 \cos(\theta) + x_2 \sin(\theta) = t\} \tag{2.6}$$

Since all lines crossing the unit sphere can be parametrised as above, inserting Eq. 2.6 into Eq. 2.4 results in a linear mapping $\mathcal{R}$ defined as

$$f(\theta, t) = \mathcal{R}u(\theta, t) = \int_{\ell(\theta,t)} u(x)\mathrm{d}x \tag{2.7}$$

which is the Radon transform. During a CT scan, for each fixed emission angle θ the position t is varied. This results in the so-called parallel-projection, which is depicted in Figure 2.3b. In this case, the Radon transform is described formally as

$$\mathcal{R}u(\theta, t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u(x)\delta(x_1 \cos\theta + x_2 \sin\theta - t)\, dx \tag{2.8}$$

where $\mathcal{R}u(\theta, t)$ is a Radon projection of the object with attenuation $u(x)$. Furthermore, t is the position of the detector bin, θ is the angular position of the X-ray tube and $\delta(\cdot)$ is defined to be the Dirac-delta function.

2.3 Ill-posedness of the Radon inverse

When one aims to reconstruct an object $u(x)$ from the given measurement $f(\theta, t)$, a problem arises regarding the inverse of the Radon operator $\mathcal{R}$. Due to the unboundedness of the inverse Radon transform [18], reconstruction becomes an ill-posed inverse problem. Any inverse problem can be mathematically described as

$$y = \mathcal{A}x \tag{2.9}$$

where the unknown $x \in X$ is the solution of the equation above, $y \in Y$ the known noiseless measurement and $\mathcal{A}$ the operator defined as $\mathcal{A} : X \to Y$, where both X and Y are Banach spaces. An inverse problem is ill-posed, if it is not well-posed. The latter is defined according to Hadamard [19] as follows.

Definition 1 *Let $\mathcal{A} : X \to Y$ be a forward mapping. Then the inverse problem $y = \mathcal{A}x$ is well-posed if all of the following points hold:*

- *a solution exists*
- *the solution is unique*
- *the solution's behaviour continuously depends on the measured data*

Ofttimes solutions do not fulfill the third requirement due to modelling and measurement errors. This is the result of the operator $\mathcal{A}$ being just a model for the underlying process. When noise is incorporated, the model is defined as

$$y^\delta = \mathcal{A}x + \epsilon \text{ or } y^\delta = \mathcal{A}x^\delta \tag{2.10}$$

with ϵ being the measurement error of a certain unknown distribution with $\|\epsilon\| \leq \delta$ and $\|y^\delta - y\| \leq \epsilon$. The ill-posedness also occurs in the inverse Radon transform $\mathcal{R}$, which cannot be defined. Due to the unboundedness of the inverse Radon transform [18], the problem of reconstructing an image x from a given sinogram y by means of the inverse of $\mathcal{R}$ is an ill-posed inverse problem. The unboundedness of $\mathcal{R}$ results in the fact that small changes in measurements might lead to significant differences in reconstructions. As a consequence, we cannot just use the inverse of the forward operator. To work around this issue, specific reconstruction algorithms for CT have been developed that estimate or learn the Radon inverse operator. Several of these algorithms that are currently used in practice are described below.

2.4 CT image reconstruction

One of the solutions for accurate image reconstruction is to use an approximation (reconstruction technique) for the inverse Radon transform that is well-posed. A standard method in CT that is used for image reconstruction is filtered back-projection (FBP) [20]. This method is an analytic reconstruction algorithm, that reconstructs image slices from projection data and applies a convolution filter in Fourier domain to remove blurring. FBP is an efficient algorithm, however it is infeasible when the projection data is noisy. This is due to the fact that in FBP it is assumed that X-rays travel along a straight line and that the X-ray source is an infinitely small focal spot [21], which results in enhancement of the noise of the projection data.

Within several decades iterative recontruction (IR) algorithms were preferred over FBP, since they are able to handle low-dose, hence noisy, CT data. These algorithms can be split in two categories, namely hybrid IR and model-based IR. The first type of IR filters the projection data iteratively to reduce any artifacts, after which the data is back-projected resulting in a reconstruction. This reconstruction is filtered as well in the image domain to reduce image noise. The second type of IR back-projects the data onto the image domain and directly afterwards, a forward projection is performed on the same image. The true and artificial projection data are then compared to update the image reconstruction. Several examples of hybrid IR algorithms are ASIR (adaptive statistical iterative reconstruction) and SAFIRE (sinogram-affirmed iterative reconstruction)[22][23]. Other examples of model-based IR algorithms are ASIR-V and ADMIRE (advanced modeled iterative reconstruction) [24][25]. Even though the reconstruction time is either minimal or average in all algorithms, the advantages of these types of algorithms is their strength in artifact reduction and noise reduction, where the latter is performed exceptionally well [5]. A reason to consider other types of reconstruction techniques, is the fact that IR algorithms rely on manually designed prior functions. Deep learning techniques on the other hand overcome this issue by learning the prior features.

Many deep learning reconstruction techniques have emerged in medical imaging which outperform iterative methods [26]. The strength of deep learning methods is that images can be reconstructed from data that is noisy and of poor quality. The deep learning techniques learn a reconstructor that is an approximation of (2.8) without any prior information. The earliest deep learning methods are based on convolutional neural networks (CNNs) [27] with an encoder-decoder structure. The DEAR-3D network [28] uses the structure of an optimized Wasserstein generative adversarial network (WGAN) [29] to perform reconstructions. This method has proven to perform better than the other CNN-type networks and the conventional FBP. In general, most recent deep learning methods continued using the CNN or GAN structure of the reconstruction task. All mentioned networks operate in the image domain,

where a reconstruction is retrieved using a regular FBP algorithm, after which image quality is enhanced using a deep neural network. An example of a deep learning method that operates in projection (sinogram) domain is the ADAPTIVE-NET, which is a model that first filters the sinogram using convolutional layers after which the sinogram is projected onto the image domain with a reconstruction as results [30].

2.5 Semantic segmentation in medical imaging

When reconstructed images are obtained from a CT scan, medical image analysis is performed to detect the position and/or size of a pathology or anatomy in a patient's body. This is often done by means of semantic segmentation. Deep learning semantic segmentation assigns a class or category to each pixel in an image, making this a pixel-level classification technique. When all pixels are assigned to a certain category, clustered pixels are labeled as one substructure and, hence, one class. This method allows to perform precise analysis of medical images of anatomic data. Early approaches were mostly based on edge detection and matching algorithms [31][32]. The implementation of machine learning techniques started only recently, with Li et al. [33] being the first ones to use support vector machines in combination with level sets for body data segmentation. One of the major drawbacks of the described segmentation techniques is the difficulty in extracting discriminating features as a result of, for example, noise, blur and low contrast in medical images. However, this has led to major developments in deep learning methods, which do not require hand-crafted features anymore.

CNNs have gained their popularity in deep learning segmentation methods due to their remarkable performance, since they are able to easily process images with noise, blur and low contrast [9]. One of the most widely used networks is the U-Net [34], which became a benchmark method for segmentation tasks in deep learning. The U-Net combines both low- and high-resolution feature maps by means of its symmetrical structure with skip connections, which improves the information flow and preserves more spatial information. It has an encoder-decoder type of structure, where the image is first downsampled to a certain size, after which it is upsampled again to the original size. Trainable convolution kernels within the downsampling and upsampling pathway extract features from the image and transform these features into a pixel-wise classification mask. Many other segmentation networks have been based on the architecture of the U-Net, such as the 3D U-Net [35] which is fit for handling 3D medical data and provide segmentations. Another network, which has the U-Net as basis, is the V-Net [36]. This network is deeper than the U-Net, which implies that it has more downsampling units, resulting in the network having a V-shaped architecture. The V-Net is also suited for three dimensional data.

2.6 Multi-task image reconstruction and segmentation

As mentioned in Chapter 1, image reconstruction and segmentation are typically considered to be disjoint problems. Hence, it is rare to see a (deep learning) method that performs both reconstruction and segmentation. However, so-called sequential methods have been developed in various fields [37], medical imaging being one of them. Sequential methods consist of several tasks that are performed consecutively, e.g. image reconstruction followed by segmentation. One of the examples is the model proposed by Thasneem et al. [38]; the model reconstructs 3D CT images of the human head from a selected amount of slices, after which the slices are segmented to determine the anatomy of the human head. Finally, the slices are interpolated to reconstruct the missing slices.

In magnetic resonance imaging (MRI), Fourier inversion (image reconstruction in MRI) and segmentation have been performed jointly, meaning that multiple tasks are executed at the same time. In order to achieve this, the SegNetMRI network [39] has been introduced. The complete network consists of an MRI reconstruction network and an MRI segmentation network, where an encoder-decoder structure is applied throughout the whole architecture. In CT, joint methods are still limited in a sense that a model is only able to perform either pre-processing or post-processing tasks jointly. Jiang et al. present a multi-scale model in their work, which is trained jointly [40]. This model performs deformable image registration, where three deep learning models are trained to perform this task accurately with three different scale levels. Other works in research on the COVID-19 pandemic perform joint image analysis on chest CT images. The work of Amyar et al. [41] is focused on classification and segmentation of lesions, where the model developed by Goncharov et al. [42] is able to identify how much lung tissue is affected by the virus and simultaneously identify whether the lesions are indeed caused by COVID-19 or a different disease. One of the few works that does perform image reconstruction and classification jointly, is the work by Wu et al. [43]. It describes an end-to-end method to detect lung nodules in low-dose chest CT images, consisting of

iterative reconstructors and a CNN for segmentation. The model is trained sequentially and jointly to be compared afterwards.

Direct methods consist of models where raw measurement data is analysed to perform post-processing tasks on. Direct methods are explored even less than joint methods, but they still form an important addition to medical imaging. The work Lee et al. [10] makes the first steps in direct detection of intracranial hemorrhage (ICH). There, an optimized convolutional neural network, the SinoNet, is proposed as a model to identify human body parts and to identify ICH in head CT scans. Another research performed by De Man et al. [44] is focused the detection and characterization of blood vessels. The presented model is able to detect vessels in sinograms and to identify the size of the vessel and its coordinates. To the best of our knowledge, there are yet no methods in CT imaging that perform semantic segmentation directly on the projection data.

All mentioned works have shown great potential when it comes to performing post-processing tasks directly. Therefore, the question arises whether a direct segmentation model could be developed that performs at least as good as an *indirect method*. Sequential, end-to-end and joint deep learning models are examples of an indirect method, where image analysis tasks are performed either sequentially or simultaneously. Figure 2.4 shows a schematic overview of the differences between an indirect and direct method. This research is focused on exploring a direct method and to evaluate its performance in comparison to an indirect method.

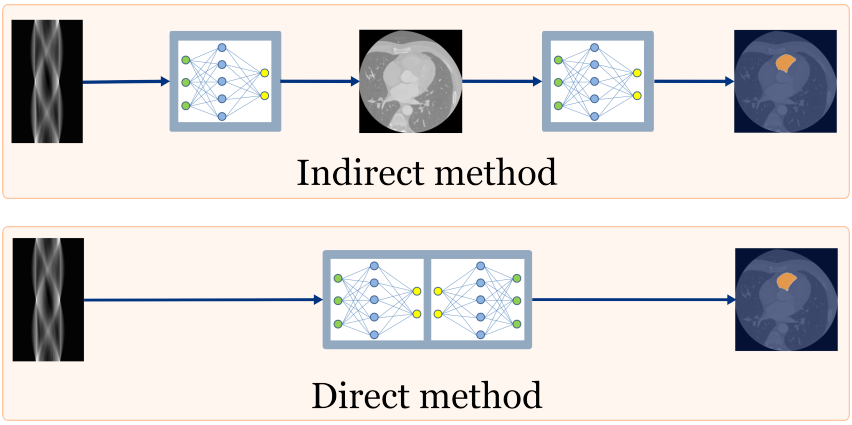


Figure 2.4: Schematic overview of an indirect method and direct method in CT imaging. In the indirect method separate methods (in this case neural networks) are used for the pre-processing and post-processing tasks. The direct method uses either combined methods or, for example, one large deep neural network to perform the post-processing task directly.

Chapter 3

Direct versus Indirect Segmentation Model

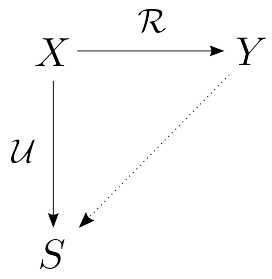
This chapter describes the mathematical model behind CT and sheds light on the direct and joint methods that are used in this research. Chapter 3.1 starts by describing the mathematical operators that occur in CT imaging and the direct operator that ought to be found in this research. Chapter 3.2 how indirect methods are expressed mathematically and how this can be put in a data-driven context. Finally, Chapter 3.3 elaborates on the geometry-based direct method that is adapted in this research, which will be compared to a joint method, that is described in Chapter 3.4.

3.1 Mathematical segmentation in CT

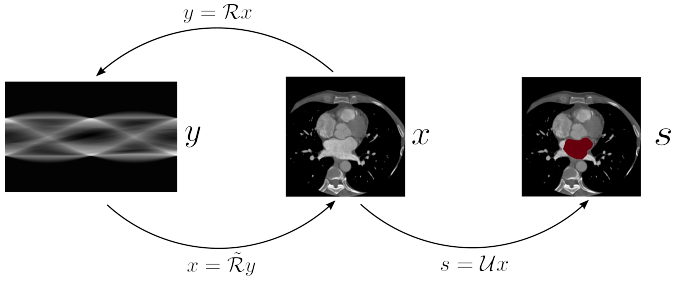
Consider the CT image reconstruction task to be an ill-posed inverse problem defined as

$$y = \mathcal{R}x + \epsilon \tag{3.1}$$

Here, $\mathcal{R}$ is defined as the forward Radon transform operator $\mathcal{R} : X \rightarrow Y$, with Y being the measurement data space and X the reconstruction space, where both are Banach spaces. Consider $y \in Y$ to be the measurement data, $x \in X$ the corresponding ground truth image and define $\tilde{x} \in X$ to be the reconstruction corresponding to image x. Furthermore, ϵ is additive noise in the projection domain with an unknown distribution. Finally, assume the inverse of the operator $\mathcal{R}$ is estimated by $\tilde{\mathcal{R}} : Y \rightarrow X$. In context of CT, y is a sinogram with a corresponding reconstructed image x. Figure 3.1b provides a visualisation of how the operators are applied on CT data.



(a) Mappings of data X with the pre-processing and post-processing operators $\mathcal{R}$ and $\mathcal{U}$, respectively.



(b) Visualisation of the application of operators $\mathcal{R}$ and $\mathcal{U}$ on image $x \in X$ and application of $\tilde{\mathcal{R}}$ on sinogram $y \in Y$.

Figure 3.1: Reconstruction and segmentation operators in CT visualised.

Next, consider the post-processing operator $\mathcal{U} : X \rightarrow S$ with a post-processing feature (segmentation mask) $s \in S$ such that

$$s = \mathcal{U}x, \tag{3.2}$$

where $\mathcal{U}$ is the forward operator $\mathcal{U} : X \to S$ and $s \in S$ the post-processing result of (a noisy) reconstruction $x \in X$. In terms of CT, the post-processing operator $\mathcal{U}$ is considered to be a semantic segmentation. Note that the operator $\mathcal{U}$ does not have an inverse, since it is non-injective for semantic segmentation. Next to this argument, estimating an operator for the inverse of $\mathcal{U}$ does not occur in practical applications. The functioning of the segmentation operator is visualised in Figure 3.1b as well. Figure 3.1a depicts a diagram of the described operators and how they are applied on image data X.

When the solution of Eq. 3.1 is used in Eq. 3.2, the segmentation task can be mathematically expressed as

$$s = (\mathcal{U} \circ \tilde{\mathcal{R}})(y) = \mathcal{U}(\tilde{\mathcal{R}}y + \eta) \tag{3.3}$$

with η the measurement noise of an unknown distribution. In an indirect method, the pre-processing operator $\tilde{\mathcal{R}}$ (reconstruction method) is expressed as a statistical estimator and the post-processing operator $\mathcal{U}$ (segmentation method) as a decision rule. For both operators, either conventional algorithms can be used or a deep learning method.

The goal of this research is to craft a mapping shown in Figure 3.1a as the dotted line from the pre-processing Y to the post-processing result S, i.e. the operator that performs *direct segmentation*. In other words, the composition $\mathcal{U} \circ \tilde{\mathcal{R}}$ must be defined as one operator. Since it is not always possible to replace the composition of two operators by one operator, one could opt for expressing the operators as neural networks (see Chapter 3.2). Combining the networks and using them in a sequential order yields in principle one network, which is optimized using one loss function. This way, $\mathcal{U} \circ \tilde{\mathcal{R}}$ can still be expressed as one operator. The method that is used for this purpose in this work is a data-driven geometry-based model. This means that a deep learning method is used to process information from measurement data for either reconstruction or a post-processing task (segmentation in this research), which is possible since reconstructions and post-processing results are in the same image space. One of the benefits of using a deep learning method is its ability to learn the operator directly from the projection data. Next to this, a direct model only considers one error that is propagated through the model, whereas in sequential approaches the measurement error is propagated through both models. In the following sections the indirect and direct methods will be described.

3.2 Indirect methods

Indirect methods consider image reconstruction and segmentation to be two tasks that are performed separately and consecutively, as is described in Chapter 2.6. According to Adler et al. [11], when using data-driven approaches, it is necessary to express reconstruction as a statistical estimator, where a post-processing task - segmentation in this case - is expressed as a decision rule. In a data-driven context, both operators are learned by a deep learning method, which makes use of features of measurement data and reconstruction data to predict the outcome of a reconstruction and segmentation, respectively. Hence, it is assumed that the reconstruction operator $\tilde{\mathcal{R}}$ operates in the sinogram domain, meaning that it projects measurement data onto an image in the image domain. The segmentation operator $\mathcal{U}$ operates solely in the image domain.

In a data-driven method, the statistical estimator $\tilde{\mathcal{R}}_{\hat{\theta}}$ estimates the reconstruction operator $\tilde{\mathcal{R}}$ and the parameter $\hat{\theta}$ is learned by means of a pre-defined reconstruction loss. In a reconstruction problem, the loss function computes the error between the ground truth image and the reconstructed image. Parameters $\hat{\theta}$ are optimized in a deep neural network such that the loss function converges to a minimal loss. This can be expressed as

$$\hat{\theta} \in \underset{\theta \in \Theta}{\operatorname{argmin}} \, \mathbb{E}_\theta \left[\ell_X(x, \tilde{\mathcal{R}}_\theta(y)) \right], \tag{3.4}$$

where the average reconstruction loss ℓ_X is minimized. Here, x is the ground truth image and $\tilde{\mathcal{R}}_\theta(y)$ the estimated reconstruction.

For the segmentation task, the goal is to find a decision rule which maps an observation (a reconstruction in this case) onto an appropriate action (segmentation mask). This decision rule is expressed as an estimator $\mathcal{U}_{\hat{\xi}}$ of the segmentation operator $\mathcal{U}$. Just as for the reconstruction task, a segmentation loss function is used to compute the error between a ground truth mask and an estimated (predicted) mask. The parameters $\hat{\xi}$ are also optimized by a deep neural network when a data-driven method is considered. The learning problem then becomes

$$\hat{\xi} \in \underset{\xi \in \Xi}{\operatorname{argmin}} \, \mathbb{E}_\xi \left[\ell_S(s, \mathcal{U}_\xi(x)) \right], \tag{3.5}$$

with s the ground truth segmentation mask, $\mathcal{U}_\xi(x)$ the estimated segmentation mask and ℓ_S the segmentation loss.

In data-driven indirect methods, which are currently used in practice, the reconstruction and segmentation models are optimized separately. The loss functions are minimized independently of each other. This means that no information of the measurement data is propagated directly to the learning of optimal prediction of segmentation masks. Since measurement data contains useful features and information that could also be used directly for the segmentation task [10], possibilities open up for a different approach. A direct data-driven method can be derived, where a model is learning end-to-end; this means that a deep learning method could be designed that trains and learns as one whole system using only one error. The model does not need to focus on unrelated tasks, such as image reconstruction, where the main goal is actually finding an accurate segmentation mask [45].

3.3 Direct method: a geometry-based model

The data-driven direct method that is adapted for direct segmentation, is derived following the works of He et al. [13] and [14]. This method method performs reconstruction based on the geometry of measurement data and is able to learn the relationship between sinusoids in the sinogram domain and pixels in the image domain. Since segmentation is fully performed in the image domain, as a segmentation mask is in principle a classification of a pixel in an image, it is possible to learn the relation between measurement and a segmentation mask. The idea of the adapted model is to learn which pixels in the image belong to which substructure and, based on this information, define a segmentation task for substructures.

Assume the continuous Radon transform in Eq. (2.8) is discretized as

$$y = Rx, \tag{3.6}$$

where $y \in \mathbb{R}^{N \cdot M}$ is the discretized sinogram data, N the number of detector bins and M the number of rotation angles. Moreover, $x \in \mathbb{R}^P$ is the discretized image to be segmented, with P being the number of pixels in the image corresponding to the sinogram. Let $R \in \mathbb{R}^{(N \cdot M) \times P}$ be the forward projection model. Here, it is assumed that the sinogram data is noiseless. With the knowledge about Eq. (3.6), an expression for the reconstructed image is derived by means of a least-squares minimization. This gives

$$x^* = (R^T R)^{-1} R^T y, \tag{3.7}$$

with x^* being the optimal solution for the reconstruction of the least-squares minimization

$$\mathcal{L}(x) = \underset{x}{\operatorname{argmin}} \frac{1}{2} \|Rx - y\|_2^2 \tag{3.8}$$

Note that $(R^T R)^{-1} R^T$ is the Moore-Penrose pseudo-inverse [46] of R, since R is not necessarily a square matrix.

It is clear that the operation on the sinogram data in Eq. (3.7) consists of two parts, namely $(R^T R)^{-1}$, which can be seen as filtering of the measurement data, and R^T which is just the back-projection matrix. He et al. [13] use these derivations to construct a three-step deep neural network, in which the sinogram data is first filtered, then back-projected onto the image domain and then processed in the image domain. The same approach can be used for segmentation of images; sinogram data can be filtered in such way that only important information for segmentation is preserved. Then, the sinogram data containing this information is back-projected onto the image domain providing a segmentation of a medical image instead of the image itself. The post-processing part remains as it is. All three steps that are mentioned are learnable by a deep neural network, which makes this network end-to-end due to only one type of loss being propagated through all three consecutive networks.

The filtering operation in the deep neural network is expressed as follows:

$$\hat{y}(k, m) = \tanh\left(\sum_{n=1}^{N} \eta_{kn}\, y(n, m) \right) \tag{3.9}$$

Here, η denotes a parameter to be learned, k and n are the detector bin indices and m the index of the rotation angle. Mathematically, the filtering layer in the network will take up the form of a $k \times m$-sized matrix, where $k \in \{1, \ldots, N\}$ and $m \in \{1, \ldots, M\}$.

After the filtering step, the filtered sinogram $\hat{y}$ must be back-projected from the measurement domain onto the image domain. This can be considered as a matrix multiplication between the back-projection matrix R^T and the sinogram $\hat{y}$. Every sinusoid in the sinogram corresponds to a certain position (pixel) (i, j) in the image domain. Using this and the knowledge about the geometry of the sinograms, the following operation is constructed and used in the deep neural network to determine which pixel corresponds to which sinusoidal wave:

$$s(i,j) = \sum_{m=1}^{M} \gamma_{ijm} \cdot \hat{y}(n,m)|_{n=INT[i\cos(\theta_m)+j\sin(\theta_m)]} \tag{3.10}$$

Here, γ_{ijm} is a parameter to be learned, i and j are the indices in the image domain, θ_m the rotation angle corresponding to the m^{th} vector in the projection data, and INT the nearest neighbor interpolation. In nearest neighbor interpolation, the value of a random point in a pixel is defined by the nearest point. This implies that $n = INT[i\cos(\theta_m) + j\sin(\theta_m)]$ is a sinusoid in the sinogram corresponding to a certain pixel in (i, j) in the segmentation mask. To improve the quality of the segmentation, an upsampling model is added sequentially to the deep neural network. The direct segmentation model is optimized using a segmentation loss ℓ_S.

3.4 Joint model

For comparison in performance, a joint model is implemented as well. This model performs multi-task learning, meaning the model learns to reconstruct images and to perform segmentation in parallel. The joint model uses the same three layers in the model, which are all expressed as the same deep neural networks. The main difference with the direct segmentation model is that the right hand side of Eq. 3.10 does not only yield a pixel of a segmentation mask corresponding to a sinusoidal wave, but is also used to determine which sinusoidal wave corresponds to pixel $x(i, j)$ in a reconstruction. In conclusion, Eq. 3.10 is computed twice. Figure 3.2 provides a schematic overview of the joint approach.

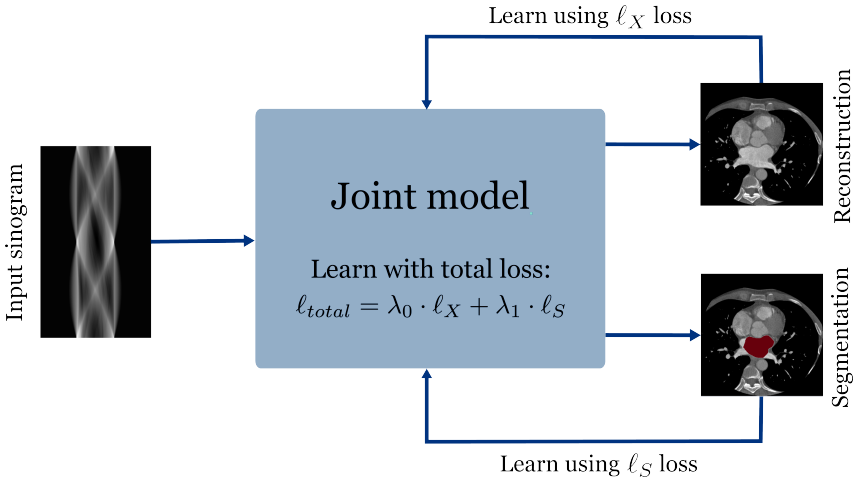


Figure 3.2: Schematic overview of the joint model.

The training of the joint model is as follows. Since the model provides segmentation masks and a reconstruction as output, two losses are computed separately for the model to train. The losses are computed as

$$\ell_{total} = \lambda_0 \cdot \ell_X + \lambda_1 \cdot \ell_S \tag{3.11}$$

with ℓ_X and ℓ_S being the reconstruction and segmentation loss respectively. Next to this, $\lambda_0, \lambda_1 \in \mathbb{R}$ are parameters that are dependent on the difference in values of ℓ_X and ℓ_S. Note that when $\lambda_1 = 0$, the joint model is equivalent to the regular model as described in [14]. When $\lambda_0 = 0$, the joint model is equal to the direct segmentation model.

Chapter 4

Materials and Methods

This chapter describes the direct-DSigNet model that is adapted as a direct method to perform segmentation using sinograms as input. Chapter 4.1 provides a detailed architecture of the direct-DSigNet, the model that is used for direct segmentation. In addition, the joint DSigNet model is described, to which the direct-DSigNet will be compared. Chapter 4.3 describes the dataset that is used for experiments. Finally, the evaluation metrics that were used to assess the performance of the network are described in Chapter 4.4.

4.1 Direct-DSigNet model

The proposed model used to perform direct segmentation is based on the downsampled-imaging-geometry-based network (DSigNet) described by He et al. [14]. The DSigNet consists of three main parts; first, the input sinograms are passed through a filtering network, which downscales feature maps in the measurement domain and transforms the geometric relationship between the sinogram and the image into a virtual relationship. After this, the embedding features of the sinogram are back-projected in the virtual back-projection network onto semantic representations of the reconstructed image using the known imaging geometry. Finally, the resulting feature maps are upsampled in the image filtering network, resulting in an image with dimensions that correspond to the original sinogram and its geometry.

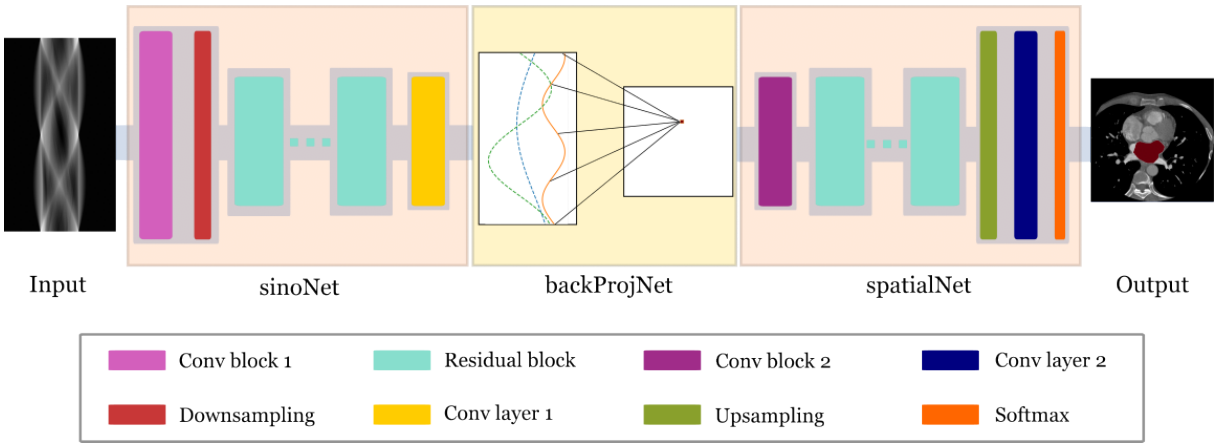


Figure 4.1: Architecture of the direct-DSigNet consisting of a sinogram filtering network (sinoNet), a virtual back-projection network (backProjNet) and an image filtering network (spatialNet). The image as output is included for visual interpretation.

The DSigNet as described in [14] is adjusted accordingly to perform direct segmentation on the sinogram data, which results in the *direct-DSigNet*. The three-sequential-networks structure is preserved in the direct-DSigNet. However the output of the direct model is the segmentation of an image corresponding to the input sinogram instead of the reconstructed image. Minor adjustments have been made in the network architecture to achieve the goal of

direct segmentation. Below, the details of the architecture and the adjustments are described. Figure 4.1 provides a visualisation of the network, where all layers and blocks of the network are visualised.

The sinogram filtering network (sinoNet) consists of a convolution block, followed by a downsampling block and finally several subsequent residual blocks. It is assumed that the input data is of size $1 \times 1152 \times 736$. The first convolution block contains 16 filtering kernels and is built from a convolution layer, followed by a normalization layer and a leaky rectified linear unit layer. After this block, the sinogram is downscaled to size $1 \times 576 \times 368$ and passed through six residual blocks. These blocks consist of three consecutive layers as in the first convolution block. Finally, the sinoNet contains a convolution layer at the end. The output of this network is a filtered sinogram of spatial size $4 \times 576 \times 368$.

This filtered sinogram is used as input for the virtual back-projection network (backProjNet). In principle, this network performs the back-projection operation from the measurement domain to the image domain. Each sinusoidal wave in the sinogram is considered and the corresponding pixels are derived according to Eq. 3.10. The network contains only one important parameter γ_{ijm} to be learned from Eq. 3.10, which results in a very large sparse matrix containing information about the imaging geometry. The output of the backProjNet are feature maps in the image domain of size $4 \times 256 \times 256$. These feature maps contain information on the segmentation masks for the corresponding image.

Finally, the feature maps are passed through the image filtering network (spatialNet), which is similar to the sinoNet. The main differences are that fewer residual blocks are used and these are succeeded by an upsampling block with a scaling factor 2. The segmentations are thus upsampled from a size of $4 \times 256 \times 256$ to $1 \times 512 \times 512$. Also, a softmax activation layer is added at the end of the network in order to differentiate between the background and pre-defined substructures. When only one substructure must be segmented, a sigmoid activation layer is used instead.

The main differences between the regular DSigNet and the direct-DSigNet are the sigmoid/softmax layer that is added in the last block and the number of output channels. The DSigNet has only 1 output channel, whereas the direct-DSigNet has 8 output channels. The number of output channels can be changed accordingly to the number of substructures that need to be segmented.

4.2 Joint DSigNet model

The network architecture of the joint model is equivalent to the architecture of the direct-DSigNet. The only difference is the number of output channels, since the joint DSigNet has 9 instead of 8: next to the segmentation masks, the model also outputs a reconstruction of an image. By analysing the joint DSigNet model and using this as a comparison for the direct-DSigNet model, one is able to see what the influence is of learning a reconstruction together with a segmentation.

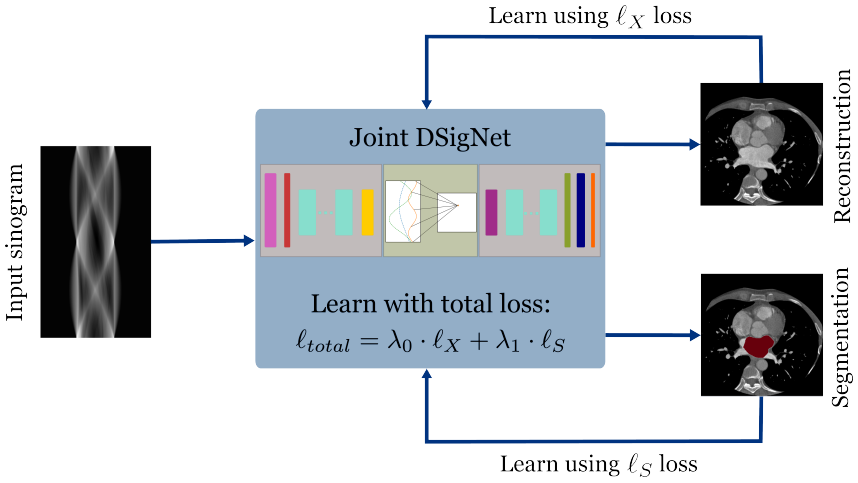


Figure 4.2: Schematic overview of the joint DSigNet model, where the architecture is preserved within the model. The total loss is computed using the weighted reconstruction and segmentation loss and is propagated end-to-end through the whole network.

4.3 Dataset and preprocessing

The public dataset of Multi-Modality Whole Heart Segmentation (MM-WHS) [47] was used in order to train and test the model. The MM-WHS dataset contains 120 images, of which 60 are cardiac CT/CTA scans in 3D, and covers all substructures of the heart from the upper abdominal to the aortic arch. The slices of the scans were acquired in the axial view and have an inplane resolution of 0.78×0.78 mm. The average slice thickness is 1.60 mm. The dataset consisted of the images and corresponding segmentations of seven whole heart substructures, whose corresponding original (as defined in [47]) and new labels are shown in Table B.1 in Appendix B. The labels were changed to be correctly read by the direct-DSigNet model. A schematic image of a heart is included in Figure 4.3 to show where each substructure is located in the heart.

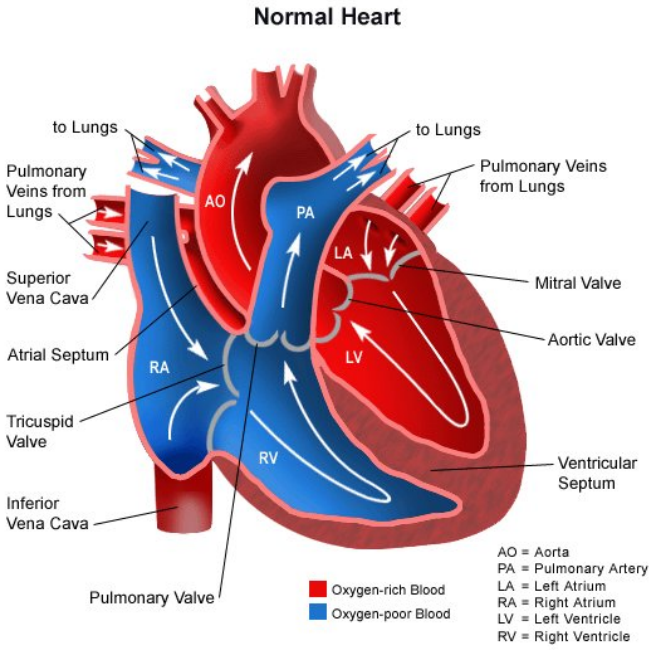


Figure 4.3: Anterior cross-sectional anatomy of the heart [48]. The schematic image shows various layers of the heart, excluding the myocardium. The myocardium is the cardiac muscle right from and below the left ventricle. The segmentation masks in the MM-WHS dataset contain the following substructures: myocardium, left and right atrium, left and right ventricle, ascending aorta and pulmonary artery.

The selected subset of CT images contains 20 images and was split in 18 training images, 1 validation image and 1 test image. The provided CT test dataset was not used, since it did not contain segmentations of the substructures for evaluation. The selected subset of the MM-WHS data contained only images and corresponding masks, but no sinograms. To generate sinograms, two generation methods are chosen, namely the LoDoPaB technical pipeline [49] and the method provided in the work of He et al. [14].

The LoDoPaB technical reference is a framework that generates low-dose CT measurement data and corresponding reconstructions. This framework can also be used to generate measurement data without noise. Code provided in the technical reference makes use of the ASTRA toolbox [50] and the Operator Discretization Library (ODL) [51]. These libraries are used to generate measurement data using certain geometries. The geometry of the LoDoPaB technical reference to generate the dataset is given in Appendix A.1. There must be noted that the distance metrics were unknown, hence have been acquired by trial-and-error. When generating the dataset, the images of the MM-WHS dataset are downsized from a resolution of 512 px $\times$ 512 px to a resolution of 362 px $\times$ 362 px. The image domain is of size 26 cm $\times$ 26 cm. It is assumed that measurements were made with 513 equidistant detector bins and from 1000 equidistant angular positions θ where $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. The input shape of the tensors for the DSigNet has been adjusted to $1 \times 1000 \times 513$ and the output shape of the tensors to $1 \times 362 \times 362$.

The other method of generating sinograms and corresponding reconstructions is the method provided in [14]. There, measurement data is acquired using only the ASTRA toolbox. The geometry used for data generation is given in

Appendix A.2. There, the number of detector units is set to 736 and the number of angular positions over 180 degrees is 1152. Furthermore, it is assumed that the images and segmentations are of size 512 px $\times$ 512 px, whereas the sinograms are 1152 px $\times$ 736 px. The architecture of the DSigNet did not need to be adjusted.

4.4 Evaluation and performance metrics

First, the DSigNet is assessed on reconstruction, to determine whether it is able to process the generated sinograms. To measure the performance and optimize the DSigNet, the L_1-loss is used, which is defined as

$$L_1 = \sum_{i=1}^{P} |p_i - \hat{p}_i| \tag{4.1}$$

with P the total number of pixels, p_i the true value of a pixel i in the image and $\hat{p}_i$ the estimated value. The L_1-loss minimizes the sum of all absolute differences between true and estimated values of the pixels. This loss type is preferred because of its robustness. Next to this, the neural network is able to make more smooth predictions when training with the L_1-loss than with any other loss function. This loss is suited to optimize a model learning on large sparse datasets [52].

The reconstruction performance of the regular and joint DSigNet will be assessed by determining the peak signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM). PSNR is a quality measure used to quantify reconstruction quality. It is expressed on the logarithmic scale and the quantity is measured on the decibel scale. Typical values for PSNR lie between 30 dB and 50 dB after reconstruction [53]. PSNR values below 20 dB indicate that a reconstruction is too noisy and a lot of image quality is lost. PSNR is defined as

$$PSNR = 20 \times \log_{10} \left(\frac{MAX_{I_O}}{\sqrt{MSE}} \right), \tag{4.2}$$

with MAX_{I_O} the maximal pixel value (intensity) of the original image and the MSE the mean-squared error defined as

$$MSE = \frac{1}{P} \sum_{i=0}^{m} \sum_{j=0}^{n} \left(I_O(i,j) - I_R(i,j) \right)^2 \tag{4.3}$$

Here, $P = m \times n$ is the total number of pixels, I_O and I_R are the pixel values of the original and reconstructed image, respectively. SSIM is a measure that quantifies the similarity between the original and reconstructed image. SSIM values range between -1 and +1, where +1 implies that the reconstructed image is identical to the original image and -1 implies that there is no similarity between both images. SSIM is defined as

$$SSIM = \frac{(2\mu_O\mu_R + c_1)(2\sigma_{OR} + c_2)}{(\mu_O^2 + \mu_R^2 + c_1)(\sigma_O^2 + \sigma_R^2 + c_2)}, \tag{4.4}$$

with μ_O and μ_R being the luminance (mean of all pixel intensities) of the original and reconstructed image, respectively, σ_O and σ_R the contrast (standard deviation of all pixel intensities) of the original and reconstructed image, respectively, and c_1 and c_2 the division stabilizers defined by $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$ with $L = 2^b - 1$ a dynamic range for pixel values. There, b is the total number of bits per pixel and k_1 and k_2 constants.

To assess the performance for semantic segmentation, the Dice similarity coefficient is used. Assume p_i and q_i are corresponding pixel values in a ground truth image and predicted image, respectively. Note, that this is also applied to segmentation masks. The Dice coefficient is then defined as follows

$$DSC = \frac{2 \sum_{i=1}^{P} p_i q_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} q_i^2} \tag{4.5}$$

with N being the total number of pixels in both images (or segmentation masks) and considering the denominator cannot be 0. Usually the values of p_i and q_i take up values 0 or 1. This results in the Dice similarity coefficient score falling in the interval $DSC \in [0, 1]$. Putting this in the scope of this research, the Dice similarity coefficient can be viewed as the ratio between twice the intersection of the segmentation mask and both segmentation masks added

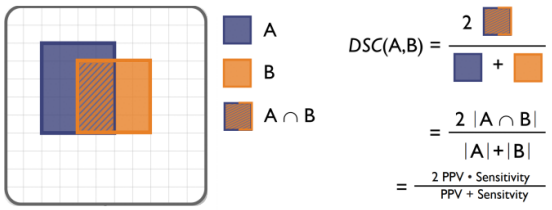


Figure 4.4: Visualisation of the Dice similarity coefficient, obtained from the work of Reinke et al. [54].

(see Figure 4.4). Note that if there is no overlap between both segmentation masks, the Dice similarity coefficient will equal 0.

Since the total number of pixels in both types of segmentation masks are considered globally (over the whole image) and locally (considering a single mask), the accuracy will be positively influenced by the Dice loss. The Dice loss is defined as $1 - DSC$ and is used to optimize the direct-DSigNet.

Chapter 5

Experiments and Results

This chapter describes performed experiments to assess the regular, joint and direct-DSigNet. First, in Chapter 5.2 two pre-processing methods are compared to each other to see how the regular DSigNet processes the datasets. Next, the segmentation results of the joint and direct-DSigNet are compared to each other in Chapter 5.3. Lastly, the regular and joint DSigNet are compared to evaluate their performance on reconstruction to draw a conclusion on the performance of the joint DSigNet. This is described in Chapter 5.4.

5.1 Implementation details

All models have been implemented using PyTorch [55]. The MM-WHS dataset has been pre-processed using MONAI [56], which is the open-source PyTorch-based framework used for medical image analysis by means of deep learning. The dataset is augmented by adjusting the contrast of the images randomly. The models are trained using the Adam optimizer [57] with a learning rate of 10^{-3}, a training batch size of 2 and a validation batch size of 10. All models are trained for 2000 epochs, with a checkpoint at 500, 1000, 1500 and 2000 epochs. The training of all models has been done on an NVIDIA Quadro RTX 6000 24GB GPU.

5.2 Image reconstruction for different geometries

Before training the direct-DSigNet model to predict segmentation masks, it is important to see how the original DSigNet model behaves when the MM-WHS dataset is used as input. Due to the lack of measurement data and corresponding reconstructions, the dataset has been generated using the LoDoPaB pipeline geometry and the geometry of the DSigNet pipeline as described in Chapter 4.3. Due to the difference in geometry, measurement data size and reconstruction size, it is essential to establish whether the DSigNet is able to process either of the generated dataset or both. To establish which generation method is more suitable, two datasets are generated on which the regular DSigNet is trained. The validation results are shown in Figure 5.1.

It becomes clear from the validation results in Figure 5.1 that the difference between the two generation methods is large. When looking at the LoDoPaB generated results, the reconstructed image that results from 500 epochs of training is blurred and the size of the heart is different than the size in the ground truth image. Training for more epochs does not results in better reconstructions; the grey area representing the heart is diminishing, however, the image is still blurred and heart substructures are not visualised. Also, the boundaries of the heart are not smooth nor clear. The results for the reconstructions with the DSigNet geometry on the other hand do improve. Background noise is present in all four reconstructions, but decreases slightly. Yet, the reconstruction after 2000 epochs contains more blur than the reconstruction after 1500 epochs.

Since the validation results of the DSigNet on the dataset with DSigNet geometry provides more accurate reconstructions than the dataset with LoDoPaB geometry, all further experiments have been performed using the first-mentioned dataset.

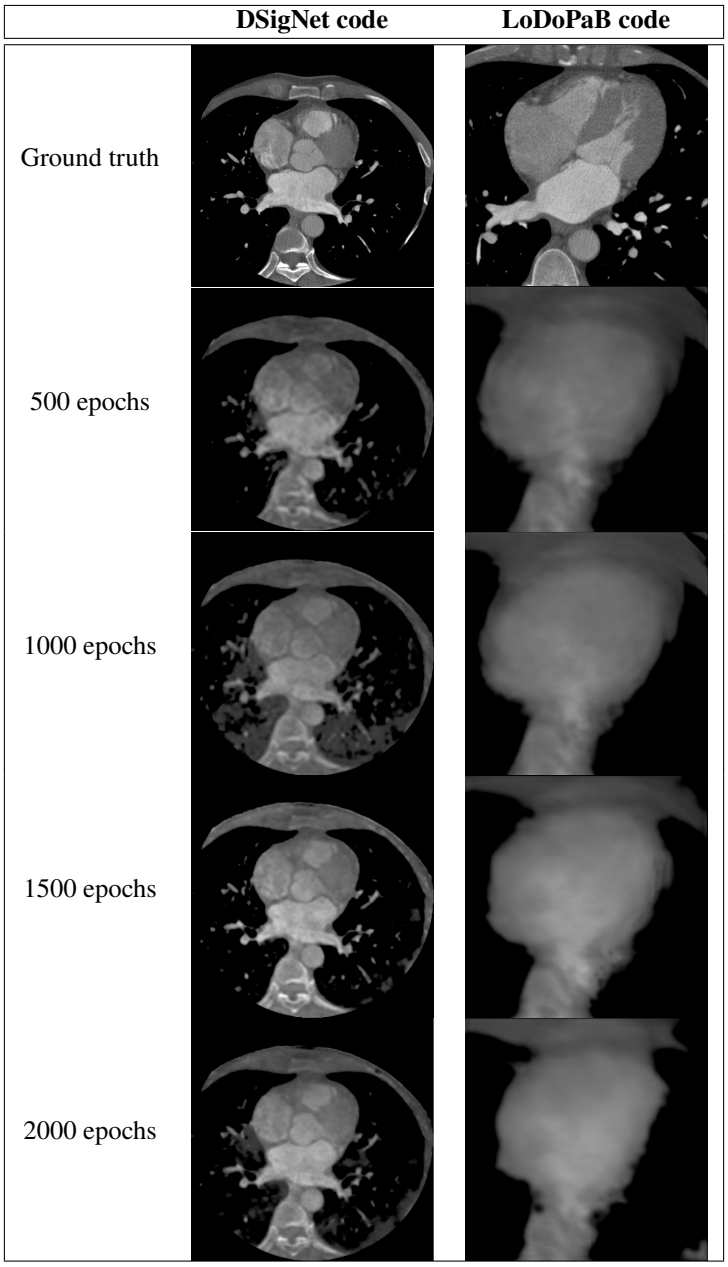
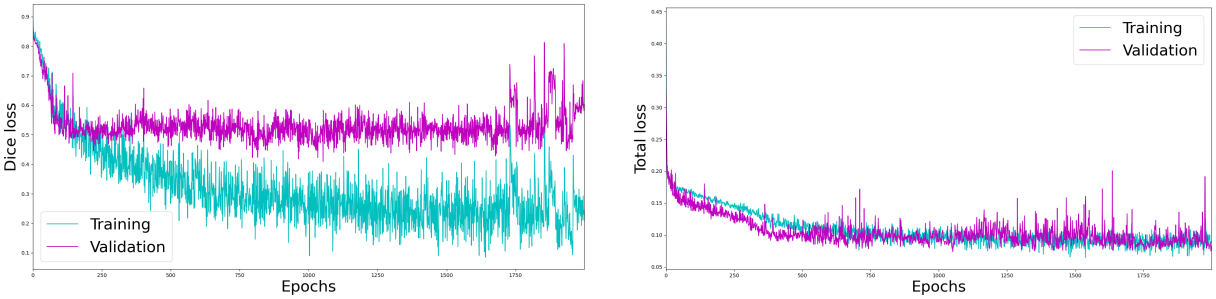


Figure 5.1: Validation reconstruction results for the DSigNet with data being generated by the LoDoPaB technical pipeline and the DSigNet code [14].

5.3 Joint and direct semantic segmentation

To assess the performance of the direct-DSigNet model for segmentation, the model is compared to the joint DSigNet, which performs reconstruction and segmentation simultaneously. The direct-DSigNet is trained using the Dice loss, whereas the joint DSigNet is trained using the total loss as defined in Eq. 3.11, where the parameters are fixed as $\lambda_0 = 0.9$ and $\lambda_1 = 0.1$. The training and validation losses are depicted in Figure 5.2b. Furthermore, the Dice coefficient score for each heart substructure is computed for both models after testing. The results are supported by Figure 5.3, which shows the results of predicted segmentation masks after testing the joint DSigNet model and the direct-DSigNet.

It is clear from Figure 5.2 that the direct-DSigNet has a less stable learning curve than the joint DSigNet, which is visible in the training and validation curves of both networks in Figure 5.2. Figure 5.2a shows that the training and validation curve of the direct-DSigNet is fluctuating a lot around a weighted average, whereas the training and



(a) Direct segmentation.

(b) Joint segmentation.

Figure 5.2: Training and validation results of the direct-DSigNet (a) and joint DSigNet (b). The first model is assessed by means of the Dice loss, where the latter is assessed with the total loss as defined in Eq. 3.11.

validation curves of the joint DSigNet are more stable, as is depicted in Figure 5.2b. Besides the fluctuation, the loss of the direct model is higher than the loss of the joint model, especially during validation. Both models tend to overfit after a certain number of epochs. For the direct-DSigNet, this starts after ∼80 epochs. For the joint DSigNet this starts at ∼200 epochs and grows gradually.

Epochs	1	2	3	4	5	6	7
500	**0.2157**	0.1619	0.1904	0.0084	0.0482	0.0482	0.1186
1000	0.2129	**0.1662**	**0.2037**	**0.0090**	0.0932	0.0761	0.1514
1500	0.2039	0.1583	0.1987	0.0080	**0.0940**	**0.0881**	**0.1734**
2000	0.1967	0.1559	0.1890	0.0079	0.0861	0.0879	0.1537

(a) Direct segmentation.

Epochs	1	2	3	4	5	6	7
500	0.1752	0.1529	0.1452	**0.0161**	**0.1468**	0.0649	0.0936
1000	0.1904	**0.1571**	0.1919	0.0082	0.0905	0.0939	**0.1635**
1500	0.1930	0.1569	0.1827	0.0079	0.0786	0.0870	0.1512
2000	**0.2004**	0.1428	**0.1997**	0.0090	0.0498	**0.1059**	0.1221

(b) Joint segmentation.

Table 5.1: Segmentation test results of the direct-DSigNet (a) and joint DSigNet (b). The Dice coefficient score is given for each heart substructure as defined in Appendix B for 500, 1000, 1500 and 2000 epochs.

Table 5.1 shows the Dice coefficient score of each heart substructure after testing of the direct- and joint DSigNet. In general, the myocardium (1) has the highest score out of all the heart substructures. The right atrium (4) has the lowest score, where the scores of direct segmentation are slightly higher than the scores of joint segmentation. For each heart substructure the highest Dice coefficient score has been marked. No score is the highest for direct segmentation test results after 2000 epochs. The highest scores do vary for the joint segmentation test results.

The information on the results of Figure 5.2 and Table 5.1 can now be used to interpret the visualisations of the predicted masks in Figure 5.3. In all images one can see that most segmentation masks for joint and direct segmentation do not correspond to the ground truth masks. In testing, the segmentation masks correspond more to the ground truth image, than in validation of the direct segmentation model. Some segmentation masks are not predicted, or disappear after more epochs. For example, at 2000 epochs the right atrium (4) disappears in the mask prediction of the direct-DSigNet, while the joint DSigNet is still able to show where the mask must be located. For the joint model, the predictions improve per 500 epochs in both validation and testing. In validation the model is not able to predict the location of the aorta (6), whereas the model is able to do so in testing. It is clear that the left atrium is the easiest shape and location for both models to predict in testing after 1000 epochs. The same holds for the right atrium (2).

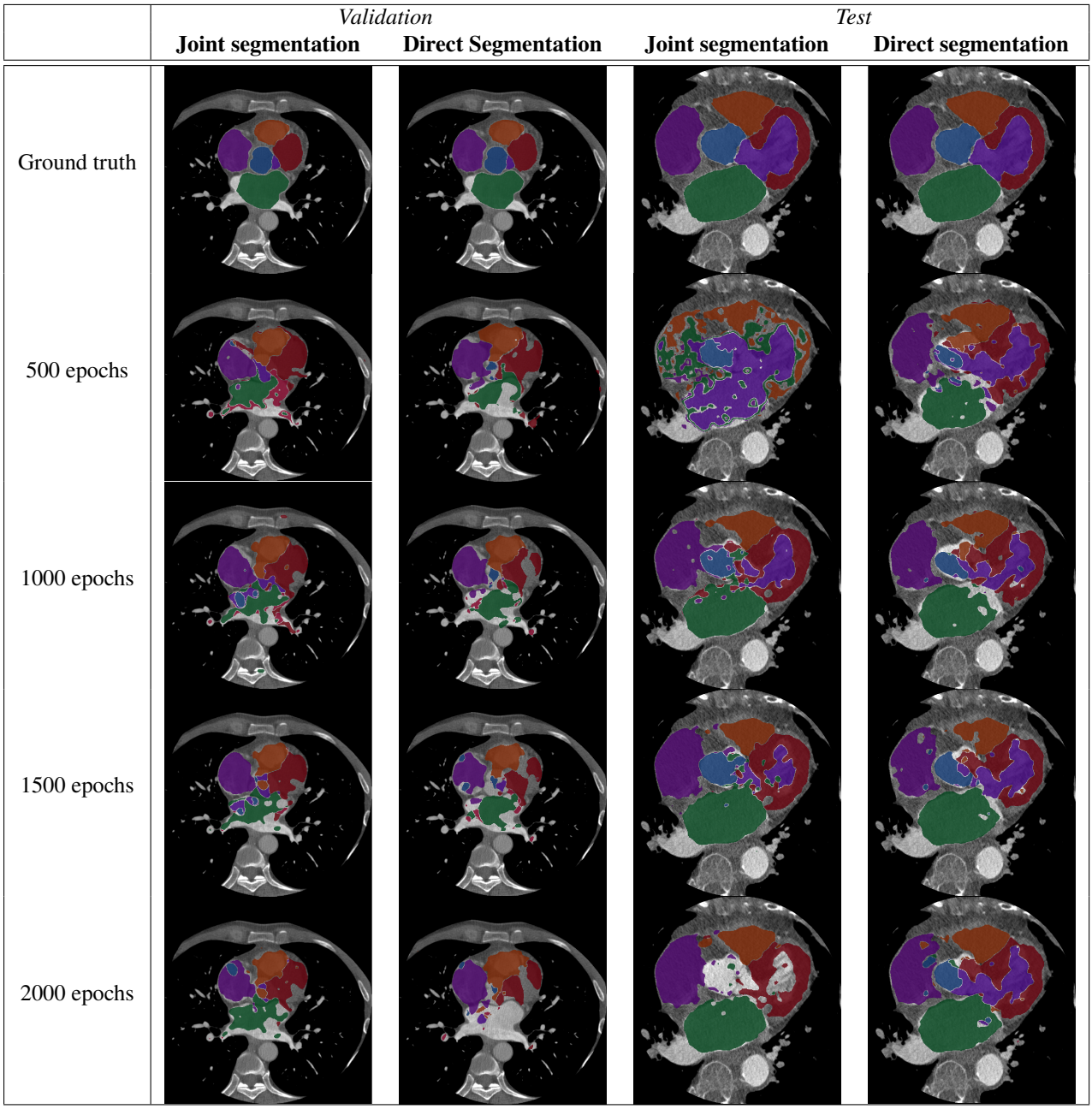
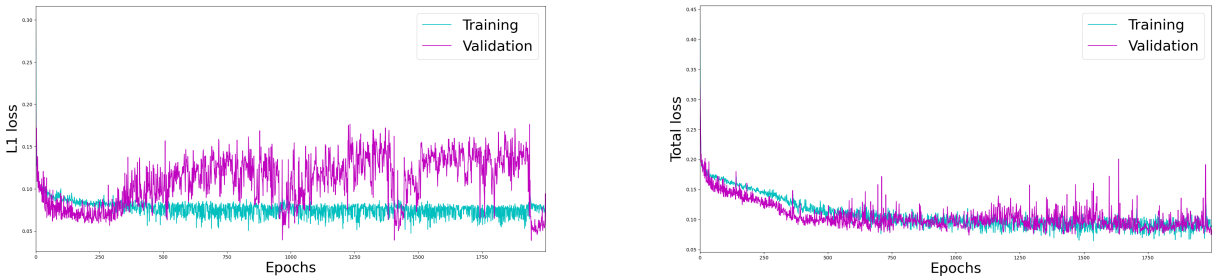


Figure 5.3: Segmentation results after validation and testing of the joint DSigNet and direct-DSigNet. The validation image is the 148^{th} slice of the 19^{th} patient, the test image is the 200^{th} slice of the 20^{th} patient. In joint training $\lambda_0 = 0.9$ and $\lambda_1 = 0.1$ are fixed. The segmentation masks are as follows: 1 - red, 2 - green, 3 - purple, 4 - magenta, 5 - orange, 6 - blue, 7 - yellow.

5.4 Joint and direct image reconstruction

Next to the segmentation results, the reconstruction results are also considered for the assessment of the joint DSigNet model. The joint DSigNet is compared to the regular DSigNet to determine whether the model performs at least as good as the conventional model. The regular DSigNet is trained using the L_1 loss, where the joint-DSigNet is trained with a loss and parameters as described in Chapter 5.3.

The loss curves in Figure 5.4b is the same as for Figure 5.2b, since the same parameters for λ_0 and λ_1 remain the same as in the segmentation experiments. Both the loss curves of the training and validation of the joint DSigNet are decreasing and do not fluctuate much. This is different for the loss curves in Figure 5.4a. The training loss of the



(a) (b)

Figure 5.4: Training and validation results of the regular DSigNet (a) and joint DSigNet (b). The first model is assessed by means of the L_1 loss, where the latter is assessed with the loss as defined in Eq. 3.11.

DSigNet is rather low and starts fluctuating after ~ 350 epochs. The validation loss has much more perturbations in the curve than the training loss, which show the overfitting in the model. After ~ 350 epochs the validation loss increases, starts to fluctuate a lot and has a sequential crescent shape from that epoch onward. Right before 2000 epochs, the loss decreases considerably.

	Joint reconstruction		**Direct reconstruction**	
Epochs	*PSNR (dB)*	*SSIM*	*PSNR (dB)*	*SSIM*
500	18.5847	0.5157	**24.8169**	0.6755
1000	20.8216	0.6586	22.3957	0.7287
1500	**21.0307**	0.6543	21.8723	**0.7371**
2000	20.6731	**0.6656**	11.3737	0.6830

Table 5.2: PSNR and SSIM values for joint and regular DSigNet images after testing both models. The quality and similarity measures are computed between each reconstruction and ground truth image after 500, 1000, 1500 and 2000 epochs.

The quality and similarity measures used to assess the joint and regular DSigNet are shown in Table 5.2, where the largest PSNR and SSIM value are marked. All measures have been computed after testing both models. In general the SSIM values are close to the value of +1, especially for the regular DSigNet. The SSIM values for joint reconstruction are lower than for direct reconstruction, however, the measures are still above 0.5, which implies that similarity of the images reconstructed by the joint model is adequate. Next to this, the SSIM increases the longer the model is trained. Regarding the PSNR values, one can observe that the PSNR of joint reconstruction increases up until testing after 1500 epochs. The PSNR value of joint reconstruction after 2000 epochs is lower. The PSNR value for direct reconstruction decreases, on the other hand. The PSNR value for the regular DSigNet starts reasonable, with 24.8169, after which it decreases significantly. The lowest PSNR value is achieved after 2000 training epochs.

The results of the training and validation losses in Figure 5.2a and the quantitative measures in Table 5.2 can be used to interpret the visual results in Figure 5.5. There, validation and test output images are compared after validation and testing of the model. The quality of the validation images do correspond to the validation losses in Figure 5.4. The quality of the images after joint reconstruction does not improve and the heart substructures are blurry. However, the boundary of the heart is sharp. In comparison, the quality of the validation images for direct reconstruction does increase. The attenuations are the most similar to the ground truth image after training for 1500 epochs. When considering the test images the results from Table 5.2 do correspond to jointly reconstructed images. The quality is improved per 500 epochs, but the images are not very similar in attenuation values to the ground truth image. It is possible to distinguish the heart substructures in the test reconstructions, where this was more difficult for validation reconstructions. The visualised test reconstructions of the regular DSigNet do not correspond with the PSNR values in Table 5.2. The heart substructures are well distinguishable, and the test image reconstructed after 2000 epochs is of better quality than the test image of joint reconstruction.

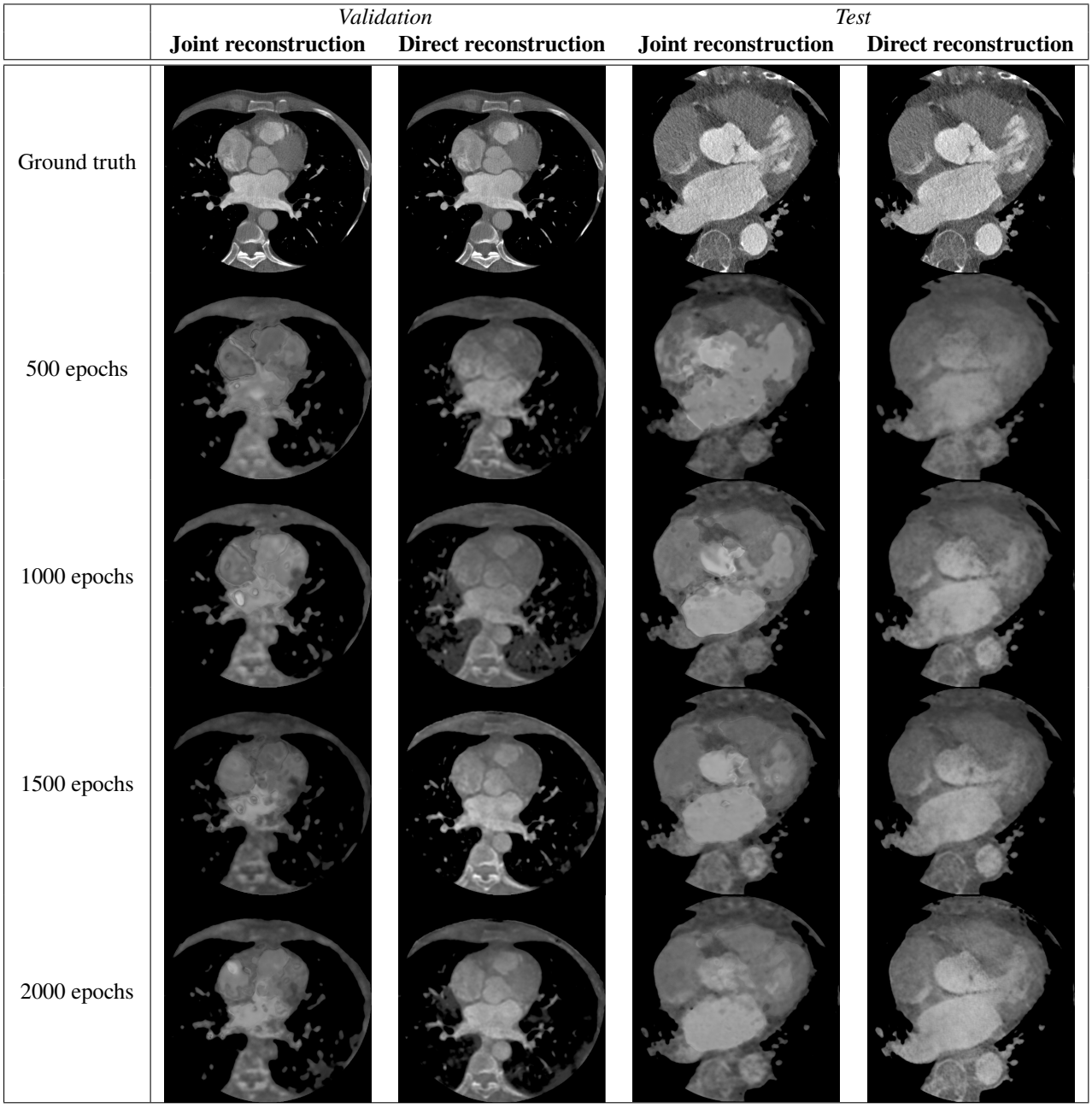


Figure 5.5: Reconstruction results after validation and testing of the joint DSigNet and regular DSigNet. The validation image is the 148^{th} slice of the 19^{th} patient, the test image is the 200^{th} slice of the 20^{th} patient. In joint training $\lambda_0 = 0.9$ and $\lambda_1 = 0.1$ are fixed.

Chapter 6

Discussion

This thesis presented a direct deep learning segmentation model, the direct-DSigNet, to evaluate its performance compared to a joint model, the joint DSigNet. The first model is able to perform direct segmentation on measurement data and to return segmentation masks as output. The latter performs segmentation and reconstruction simultaneously and provides segmentation masks and a corresponding image reconstruction. To assess the full performance of the joint DSigNet, the model is compared to the regular DSigNet to determine whether joint training will lead to more accurately predicted segmentation masks, compared to the direct-DSigNet.

First, the regular DSigNet has been trained and validated on two datasets, both being a subset of the MM-WHS dataset, but generated by two different geometries. After assessing the performance and choosing a dataset that yielded the best performance of the model, this dataset was used to perform segmentation and reconstruction experiments. The direct-DSigNet has been compared to the joint DSigNet to assess their performance on (direct) segmentation. Subsequently, the joint DSigNet has been compared to the regular DSigNet to assess the reconstruction performance of both models. Finally a conclusion is drawn regarding the performance of all models.

6.1 Reconstruction with different geometries

It appeared during the experiments that both the DSigNet and direct-DSigNet models are sensitive to the geometry of measurement data. As described in Chapter 5.2, the DSigNet model had difficulty with reconstructing an image from the input sinograms generated by the LoDoPaB technical pipeline. This dataset contained less angular positions and detector bins than the dataset generated by means of the DSigNet architecture as provided [14]. In addition, several other geometric parameters differed between the two generation methods. This is one of the main reasons for the bad performance of the DSigNet on the LoDoPaB-generated data. The generation of sinograms in the original code is based on detailed information regarding the number of voxels[1], the voxel size and information about the positioning of the scanned patient relative to the position of the X-ray source and the detector bins. This was unknown for the MM-WHS data, hence had to be estimated by trial-and-error. Since this is not feasible when one wishes to get accurate and realistic results from the experiments, the data was generated anew with the code provided for the DSigNet which did coincide with the architecture of the network and the relevant geometric information for the back-projection layer. In conclusion, correct geometry of measurements is of utmost importance for the DSigNet. This does bring up an issue, however. The ultimate goal would be to develop a model that is versatile and whose geometry parameters could be tuned. Concluding from the results in Section 5.2, the model is not able to adjust to sinograms and images which have a different geometry than the geometry defined in [14]. It is important to keep this in mind when any of the DSigNet models are explored in the future.

6.2 Direct and joint segmentation

Earlier training and validation results have shown that the performance of the direct-DSigNet is poor, therefore it makes no sense to compare the direct-DSigNet with a sequential (indirect) model. Algorithms dedicated to segmentation on the MM-WHS dataset [58] have shown to outperform the direct-DSigNet model; the worst performing

[1] Volume element of a 3D object.

algorithms achieved a Dice-loss of 0.194 ± 0.159, whereas the lowest Dice loss of the direct-DSigNet was not lower than 0.482 ± 0.059. This all, while having optimized the direct-DSigNet. Therefore, the choice has been made to compare the performance of the direct-DSigNet to a joint variant of the model, instead of an indirect model.

From the validation and test results regarding segmentation it becomes clear that the direct-DSigNet produces outputs that are not consistent with the performance metrics and the learning curves. The training and validation loss in Figure 5.2a show that the model is heavily overfitting in both training and validation, where the first-mentioned is fluctuating much more. Changing values of hyperparameters such as the learning rate did not change the behaviour of the learning curves. One of the reasons for this, is the very large number of training parameters of the direct-DSigNet in comparison to amount of data that is available [59]. The direct-DSigNet model has 39,185,869 parameters, whereas the number of available training, validation and test data is small (18 training images, 1 validation image, 1 test image). All images contain no more than 363 slices. This lack of data affects the performance of the direct-DSigNet. The joint DSigNet suffers less from overfitting, due to multiple features that must be learned. The joint model must learn features for not only reconstruction, but also segmentation.

As mentioned in Chapter 5, the Dice coefficient scores obtained during testing are not high and do not reach values above 0.2157. There can be concluded that the scores of joint segmentation are in general higher than the scores of direct segmentation when it comes to smaller substructures, whereas for the larger heart substructures the converse holds. However, when taking the outputs in Figure 5.3 into account, one can observe that the direct-DSigNet is better at locating small structures as well, such as the aorta (6). Since the test image is identical to one of the images in the training set, the perfomance of both joint and direct-DSigNet increases during testing. Keeping the two mentioned discussion points in mind, it is essential in further research that the chosen dataset does not only contain more slices and/or patient images, but is also well-balanced. This means that images in the validation and test dataset are not yet seen by the model during training.

Another thing that can be noted in testing of the direct segmentation model, is that early stopping could be used during the training of the model. Since the Dice coefficient scores do not improve after a training of 1500 epochs, and sometimes even decrease, one could consider to stop training at 1500 epochs or between 1500 and 2000 epochs. This does not necessarily hold for the joint model, however. For most heart substructures, the Dice coefficient score is decreasing immediately after 500 epochs, only for masks (1), (3) and (6) the scores do increase gradually. Considering the outputs of Figure 5.3, the joint model could be trained with only a subset of the heart substructure masks, such that it could focus on learning the features of masks whose location is not predicted well.

In this work, the segmentation masks are shown as image overlays. When keeping the ultimate goal of direct image analysis in mind, one could argue that segmentation masks can be learned in a different fashion. The model could output coordinates of the substructures in the patients body, instead of a segmentation mask. These coordinates can be learned directly from sinograms, like the segmentation masks. This could be useful in practice, especially during screening of a patient. Future work could focus on learning coordinates and/or location of a substructure in the image domain based on the measurement data in the sinogram domain.

6.3 Direct and joint reconstruction

When comparing results of reconstruction with the joint and regular DSigNet, one can observe that the performance of the regular DSigNet is better than the joint DSigNet. This is an interesting observation, since the joint DSigNet learns more features due to the two parallel tasks that it must perform. Considering the outcomes of the PSNR and SSIM values and their behaviour over training time of the direct model, one would assume that the output reconstructions would be less accurate than those for joint reconstruction. Nevertheless, Figure 5.5 shows that the reconstructions do improve over time and become more detailed. For the test image one could argue that this is, once again, the result of the test image being present in the training data set. In the validation image it is clear, however, that validating the model after 1500 epochs results in the best approximation concerning the intensity values of the pixels in the image. The performance of the regular DSigNet could be improved by increasing the batch size [60]. All in all, the reconstruction results for both models have shown to be of decent quality, since most of the PSNR values are above 20 dB.

The roles of the parameters λ_0 and λ_1 are significant in the computation of the total loss for the joint DSigNet model. These parameters influence the focus of the training on either reconstruction or segmentation. Changing the values of the parameters might lead to different outcomes for the reconstruction results, since the training and validation loss of the regular DSigNet was 10 times lower than the loss of the direct-DSigNet. Making λ_0 smaller would lead

to better results in reconstruction, but could worsen the outcome of segmentation training. Therefore, it is important in future experiments to determine what the main task of the joint DSigNet should be and adjust the values of λ_0 and λ_1 to this. In earlier experiments of this research, the values of λ_0 and λ_1 have been varied to $\lambda_0 = 0.8 \wedge \lambda_1 = 0.2$ and $\lambda_0 = 0.7 \wedge \lambda_1 = 0.3$, but this did not yield better segmentation nor reconstruction results than the fixed values in the experiments. On the other hand, a goal in further experiments could be to find the optimal balance between λ_0 and λ_1 in the total loss, to achieve both accurate reconstruction and segmentation. Current methods that perform (partially) joint learning are, for example, either focused on learning the optimal operators in variational methods [61][62] or do not consider the optimization of the scaling parameters [41]. Exploring the influence of the parameters on each other is valuable, since the losses that are multiplied to the parameters do not always yield the same loss values. For example, the Dice loss will always be between 0 and 1, whereas the L_1 or L_2 loss can have values larger than 1.

6.4 General remarks

In general, one can say that the performance of all three models is not as good as the performance of sequential models that are presented in [58]. Especially after testing the model that is trained for 2000 epochs, both models are unable to locate several heart substructures and some segmentation masks disappear. This behaviour is the result of ill-posedness that occurs in both models, especially in the direct-DSigNet. Image reconstruction is an ill-posed inverse problem as described before. This means that the forward map, the Radon transform, does not yield unique results when measurement data is limited [63][64]. In general, these issues are solved by adding a regularization term to the inverse problems [65], which prevents overfitting in and enhances the training in image reconstruction model. However, this is not the case in the presented DSigNet. The mathematical formulation of the model in Chapter 3.3, does not involve a regularization term, which makes the learning of the model unstable. When performing direct segmentation on measurement data, the non-unique solutions that result from the back-projection layer in the direct-DSigNet are propagated through the network, leading to incorrect segmentations. In practice, this is visible when certain heart substructures are recognized as other substructures, bone tissue, small arteries or alveoli, for example (see Figure 5.3). The same issue appears in image reconstruction, where one can see that certain heart substructures have the same attenuation as others, which results in blurry images.

A potential solution to the problem of ill-posedness is to add a regularization term to Eq. 3.6. This prevents the model from heavy overfitting and prevents model parameters to take up extreme values by adding a penalization term to the loss function. Next to this, it is possible to retrieve a feasible unique solution for the reconstruction task, and at the same time the segmentation task [66]. Adding a regularization term results in a robust solution that is less sensitive to noise than a model without a regularization term. Especially when low-dose CT measurement data is used as input for the DSigNet, a regularization term could still ensure accurate results and decent performance. Currently, the generated measurement data is assumed to be noiseless, which makes all variations of the DSigNet being modeled to process noiseless sinograms. In practice, measurement data does contain noise, even when the radiation dose is not reduced in a CT scan [67]. This means that small perturbations in measurement data lead to significant differences in reconstructions. In direct segmentation no reconstruction is made, but the same type of back-projection is applied in the direct-DSigNet model that could cause erroneous segmentations. Future work could be focused on exploring the role of noise in the performance of the direct segmentation model, how it affects the prediction of segmentation masks and the role of a regularizer in the optimization of the direct-DSigNet.

Chapter 7

Conclusion and Outlook

This work has explored the benefits and shortcomings of the direct-DSigNet, a deep learning model that is adapted from [14] to perform direct segmentation. For comparison, the DSigNet model has been adapted in the joint DSigNet, which performs image reconstruction and segmentation simultaneously. Next to this, the joint DSigNet model has been assessed on its reconstruction performance to complete the larger picture on the performance of the model. One of the main points that came to light was the importance of the geometry of the measurement data and its influence on the training and validation results of the DSigNet. The model is only able to process the geometry that is provided in the code corresponding to the work of He et al.[14]. This must be kept in mind when performing further experiments with the DSigNet and variations of it.

The direct-DSigNet has shown to produce more accurate segmentation results compared to the joint DSigNet. Nevertheless, the performance of both models has been found to be worse than expected. Training and validation of both models could be improved by adding regularization terms to the models and choosing a balanced large dataset. Exploring the potential of the direct-DSigNet further is encouraged, since the mentioned improvements might lead to better performance in segmentation. In general, the possibility of performing direct segmentation is shown to be feasible in this research.

This work has also explored the performance of the joint DSigNet on segmentation and reconstruction. The reconstruction results appeared to be more promising than the segmentation results when compared to the regular DSigNet and the direct-DSigNet, respectively. Once again, regularization plays a big role in the learning behaviour of the model. An interesting research question that follows from the experiments is the influence of the scaling parameters on each other, as a result of different orders of magnitude of the reconstruction and segmentation loss.

Bibliography

[1] W. A. Kalender, "X-ray computed tomography," *Physics in Medicine & Biology*, vol. 51, no. 13, p. R29, 2006.

[2] W. E. Thomas, W. E. G. Thomas, M. W. Reed, and M. G. Wyatt, *Oxford Textbook of Fundamentals of Surgery*. Oxford University Press, 2016.

[3] R. Smith-Bindman, J. Lipson, R. Marcus, K.-P. Kim, M. Mahesh, R. Gould, A. B. De González, and D. L. Miglioretti, "Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer," *Archives of internal medicine*, vol. 169, no. 22, pp. 2078–2086, 2009.

[4] G. Wang, Y. Zhang, X. Ye, and X. Mou, *Machine Learning for Tomographic Imaging*. IOP Publishing, 2019.

[5] M. J. Willemink and P. B. Noël, "The eevolution of image reconstruction for ct — from filtered back projection to artificial intelligence," *European radiology*, vol. 29, no. 5, pp. 2185–2195, 2019.

[6] A. Maier, C. Syben, T. Lasser, and C. Riess, "A gentle introduction to deep learning in medical image processing," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 86–101, 2019.

[7] G. Chartrand, P. M. Cheng, E. Vorontsov, M. Drozdzal, S. Turcotte, C. J. Pal, S. Kadoury, and A. Tang, "Deep learning: A primer for radiologists," *Radiographics*, vol. 37, no. 7, pp. 2113–2131, 2017.

[8] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[9] T. Lei, R. Wang, Y. Wan, X. Du, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey.," 2020.

[10] H. Lee, C. Huang, S. Yune, S. H. Tajmir, M. Kim, and S. Do, "Machine friendly machine learning: Interpretation of computed tomography without image reconstruction," *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.

[11] J. Adler, S. Lunz, O. Verdier, C.-B. Schönlieb, and O. Öktem, "Task adapted reconstruction for inverse problems," *Inverse Problems*, 2021.

[12] C. Chung, J. Kalpathy-Cramer, M. V. Knopp, and D. A. Jaffray, "In the era of deep learning, why reconstruct an image at all?," *Journal of the American College of Radiology*, vol. 18, no. 1, pp. 170–173, 2021.

[13] J. He, Y. Wang, and J. Ma, "Radon inversion via deep learning," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 2076–2087, 2020.

[14] J. He, S. Chen, H. Zhang, X. Tao, W. Lin, S. Zhang, D. Zeng, and J. Ma, "Downsampled imaging geometric modeling for accurate ct reconstruction via deep learning," *IEEE Transactions on Medical Imaging*, vol. 40, no. 11, pp. 2976–2985, 2021.

[15] G. Kracht, "Ct-scan." `https://www.lumc.nl/patientenzorg/praktisch/patientenfolders/CT-scan`. Accessed: 2022-08-17.

[16] "Computed tomography (ct)." `https://www.fda.gov/radiation-emitting-products/medical-x-ray-imaging/computed-tomography-ct`. Accessed: 2022-07-25.

[17] T. Van Leeuwen and C. Brune, "Lectures on inverse problems and imaging," April 2022.

[18] E. Candes, C. A. Sing-Long, and E. Bates, "Applied fourier analysis and elements of modern signal processing," February 2021.

[19] J. Hadamard, *Lectures on Cauchy's Problem in Linear Partial Differential Equations*. Courier Corporation, 2003.

[20] L. A. Shepp and B. F. Logan, "The fourier reconstruction of a head section," *IEEE Transactions on nuclear science*, vol. 21, no. 3, pp. 21–43, 1974.

[21] R. Schofield, L. King, U. Tayal, I. Castellano, J. Stirrup, F. Pontana, J. Earls, and E. Nicol, "Image reconstruction: Part 1 – understanding filtered back projection, noise and image acquisition," *Journal of cardiovascular computed tomography*, vol. 14, no. 3, pp. 219–225, 2020.

[22] P. B. Noël, A. A. Fingerle, B. Renger, D. Münzel, E. J. Rummeny, and M. Dobritz, "Initial performance characterization of a clinical noise–suppressing reconstruction algorithm for mdct," *American Journal of Roentgenology*, vol. 197, no. 6, pp. 1404–1409, 2011.

[23] H. Scheffel, P. Stolzmann, C. L. Schlett, L.-C. Engel, G. P. Major, M. Károlyi, S. Do, P. Maurovich-Horvat, and U. Hoffmann, "Coronary artery plaques: cardiac ct with model-based and adaptive-statistical iterative reconstruction technique," *European journal of radiology*, vol. 81, no. 3, pp. e363–e369, 2012.

[24] P. De Marco and D. Origgi, "New adaptive statistical iterative reconstruction asir-v: Assessment of noise performance in comparison to asir," *journal of applied clinical medical physics*, vol. 19, no. 2, pp. 275–286, 2018.

[25] S. Ellmann, F. Kammerer, T. Allmendinger, M. Hammon, R. Janka, M. Lell, M. Uder, and M. Kramer, "Advanced modeled iterative reconstruction (admire) facilitates radiation dose reduction in abdominal ct," *Academic Radiology*, vol. 25, no. 10, pp. 1277–1284, 2018.

[26] D. O. Baguer, J. Leuschner, and M. Schmidt, "Computed tomography reconstruction using deep image prior and learned reconstruction methods," *Inverse Problems*, vol. 36, no. 9, p. 094004, 2020.

[27] E. Ahishakiye, M. Bastiaan Van Gijzen, J. Tumwiine, R. Wario, and J. Obungoloch, "A survey on deep learning in medical image reconstruction," *Intelligent Medicine*, vol. 1, no. 03, pp. 118–127, 2021.

[28] H. Xie, H. Shan, and G. Wang, "Deep encoder-decoder adversarial reconstruction (dear) network for 3d ct from few-view data," *Bioengineering*, vol. 6, no. 4, p. 111, 2019.

[29] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*, pp. 214–223, PMLR, 2017.

[30] Y. Ge, T. Su, J. Zhu, X. Deng, Q. Zhang, J. Chen, Z. Hu, H. Zheng, and D. Liang, "Adaptive-net: deep computed tomography reconstruction network with analytical domain transformation knowledge," *Quantitative Imaging in Medicine and Surgery*, vol. 10, no. 2, p. 415, 2020.

[31] Z. Yu-Qian, G. Wei-Hua, C. Zhen-Cheng, T. Jing-Tian, and L. Ling-Yun, "Medical images edge detection based on mathematical morphology," in *2005 IEEE engineering in medicine and biology 27th annual conference*, pp. 6492–6495, IEEE, 2006.

[32] M. Lalonde, M. Beaulieu, and L. Gagnon, "Fast and robust optic disc detection using pyramidal decomposition and hausdorff-based template matching," *IEEE transactions on medical imaging*, vol. 20, no. 11, pp. 1193–1200, 2001.

[33] S. Li, T. Fevens, and A. Krzyżak, "A svm-based framework for autonomous volumetric medical image segmentation using hierarchical and coupled level sets," in *International Congress Series*, vol. 1268, pp. 207–212, Elsevier, 2004.

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

[35] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*, pp. 424–432, Springer, 2016.

[36] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571, IEEE, 2016.

[37] S. Springer, A. Glielmo, A. Senchukova, T. Kauppi, J. Suuronen, L. Roininen, H. Haario, and A. Hauptmann, "Reconstruction and segmentation from sparse sequential x-ray measurements of wood logs," *arXiv preprint arXiv:2206.09595*, 2022.

[38] A. H. Thasneem, M. M. Sathik, and R. Mehaboobathunnisa, "A fast segmentation and efficient slice reconstruction technique for head ct images," *Journal of Intelligent Systems*, vol. 28, no. 4, pp. 533–547, 2019.

[39] L. Sun, Z. Fan, X. Ding, Y. Huang, and J. Paisley, "Joint cs-mri reconstruction and segmentation with a unified deep network," in *International conference on information processing in medical imaging*, pp. 492–504, Springer, 2019.

[40] Z. Jiang, F.-F. Yin, Y. Ge, and L. Ren, "A multi-scale framework with unsupervised joint training of convolutional neural networks for pulmonary deformable image registration," *Physics in Medicine & Biology*, vol. 65, no. 1, p. 015011, 2020.

[41] A. Amyar, R. Modzelewski, H. Li, and S. Ruan, "Multi-task deep learning based ct imaging analysis for covid-19 pneumonia: Classification and segmentation," *Computers in Biology and Medicine*, vol. 126, p. 104037, 2020.

[42] M. Goncharov, M. Pisov, A. Shevtsov, B. Shirokikh, A. Kurmukov, I. Blokhin, V. Chernina, A. Solovev, V. Gombolevskiy, S. Morozov, *et al.*, "Ct-based covid-19 triage: Deep multitask learning improves joint identification and severity quantification," *Medical image analysis*, vol. 71, p. 102054, 2021.

[43] D. Wu, K. Kim, B. Dong, and Q. Li, "End-to-end abnormality detection in medical imaging," 2018.

[44] Q. De Man, E. Haneda, B. Claus, P. Fitzgerald, B. De Man, G. Qian, H. Shan, J. Min, M. Sabuncu, and G. Wang, "A two-dimensional feasibility study of deep learning-based feature detection and characterization directly from ct sinograms," *Medical physics*, vol. 46, no. 12, pp. e790–e800, 2019.

[45] T. Glasmachers, "Limits of end-to-end learning," in *Asian conference on machine learning*, pp. 17–32, PMLR, 2017.

[46] R. Penrose, "A generalized inverse for matrices," in *Mathematical proceedings of the Cambridge philosophical society*, vol. 51, pp. 406–413, Cambridge University Press, 1955.

[47] X. Zhuang and J. Shen, "Multi-scale patch and multi-modality atlases for whole heart segmentation of mri," *Medical image analysis*, vol. 31, pp. 77–87, 2016.

[48] R. K. Clark, *Anatomy and physiology: understanding the human body*. Jones & Bartlett Learning, 2005.

[49] J. Leuschner, M. Schmidt, D. O. Baguer, and P. Maaß, "The lodopab-ct dataset: A benchmark dataset for low-dose ct reconstruction methods," *arXiv preprint arXiv:1910.01113*, 2019.

[50] W. Van Aarle, W. J. Palenstijn, J. De Beenhouwer, T. Altantzis, S. Bals, K. J. Batenburg, and J. Sijbers, "The astra toolbox: A platform for advanced algorithm development in electron tomography," *Ultramicroscopy*, vol. 157, pp. 35–47, 2015.

[51] J. Adler, H. Kohr, A. Ringh, J. Moosmann, M. J. Ehrhardt, G. R. Lee, O. Verdier, J. Karlsson, W. J. Palenstijn, O. Öktem, *et al.*, "Odlgroup/odl: Odl 0.7. 0," *Zenodo*, 2018.

[52] K. Janocha and W. M. Czarnecki, "On loss functions for deep neural networks in classification," *arXiv preprint arXiv:1702.05659*, 2017.

[53] N. Dey, A. S. Ashour, F. Shi, and V. E. Balas, *Soft Computing Based Medical Image Analysis*. Academic Press, 2018.

[54] A. Reinke, M. Eisenmann, M. D. Tizabi, C. H. Sudre, T. Rädsch, M. Antonelli, T. Arbel, S. Bakas, M. J. Cardoso, V. Cheplygina, *et al.*, "Common limitations of image processing metrics: A picture story," *arXiv preprint arXiv:2104.05642*, 2021.

[55] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang,

J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, Curran Associates, Inc., 2019.

[56] T. M. Consortium, "Project monai," Dec. 2020.

[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[58] X. Zhuang, L. Li, C. Payer, D. Štern, M. Urschler, M. P. Heinrich, J. Oster, C. Wang, Ö. Smedby, C. Bian, *et al.*, "Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge," *Medical image analysis*, vol. 58, p. 101537, 2019.

[59] B. S. Everitt and A. Skrondal, "The cambridge dictionary of statistics," 2010.

[60] S. L. Smith, P.-J. Kindermans, C. Ying, and Q. V. Le, "Don't decay the learning rate, increase the batch size," *arXiv preprint arXiv:1711.00489*, 2017.

[61] Y. E. Boink, S. Manohar, and C. Brune, "A partially-learned algorithm for joint photo-acoustic reconstruction and segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 1, pp. 129–139, 2019.

[62] F. Lauze, Y. Quéau, and E. Plenge, "Simultaneous reconstruction and segmentation of ct scans with shadowed data," in *International Conference on Scale Space and Variational Methods in Computer Vision*, pp. 308–319, Springer, 2017.

[63] K. T. Smith, D. C. Solmon, and S. L. Wagner, "Practical and mathematical aspects of the problem of reconstructing objects from radiographs," *Bulletin of the American Mathematical Society*, vol. 83, no. 6, pp. 1227–1270, 1977.

[64] R. Estrada and B. Rubin, "Null spaces of radon transforms," *Advances in Mathematics*, vol. 290, pp. 1159–1182, 2016.

[65] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of inverse problems*, vol. 375. Springer Science & Business Media, 1996.

[66] J. Zhang, Q. He, C. Wang, H. Liao, and J. Luo, "A general framework for inverse problem solving using self-supervised deep learning: validations in ultrasound and photoacoustic image reconstruction," in *2021 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, IEEE, 2021.

[67] W. Mazrani, K. McHugh, and P. Marsden, "The radiation burden of radiological investigations," *Archives of disease in childhood*, vol. 92, no. 12, pp. 1127–1131, 2007.

Appendix A

Geometry of Measurements

Below, the geometries of measurements that were used for generation of sinograms are provided Table A.1 and A.2.

A.1 Sinogram generation with LoDoPab settings

Parameter	Value
Number of voxels	362 px $\times$ 362 px
Voxel size	0.78 mm $\times$ 0.78 mm
Number of detector bins	513
Detector bin size	1.25 mm
Number of angular positions	1000
Offset from origin	(0, 0)
Interval of angles	$[-\frac{\pi}{2}, \frac{\pi}{2}]$
Number of input slices	1
Distance source to origin (DSO)	595.0 mm
Distance origin to detector (DOD)	490.6 mm
Distance source to detector (DSD)	1085.6 mm

Table A.1: Geometry parameters used for generation of sinograms by means of the LoDoPaB technical pipeline.

A.2 Sinogram generation with DSigNet settings

Parameter	Value
Number of voxels	512 px $\times$ 512 px
Voxel size	0.6641 mm $\times$ 0.6641 mm
Number of detector bins	736
Detector bin size	1.3696 mm
Number of angular positions	1152
Offset from origin	(0, 0)
Interval of angles	$[0, \pi]$
Number of input slices	1
Distance source to origin (DSO)	595.0 mm
Distance origin to detector (DOD)	490.6 mm
Distance source to detector (DSD)	1085.6 mm

Table A.2: Geometry parameters used for generation of sinograms by means of the ASTRA toolbox.

Appendix B

Labels of the MM-WHS Dataset

For the ease of implementation, the labels of the MM-WHS dataset have been changed. Original labels are shown in the center column of Table B.1, where the new labels are given in the right column. Re-labeling masks this way is an important data pre-processing step such that the model is able to make accurate predictions.

Substructure	Original label	New label
Myocardium of left ventricle	205	1
Left atrium blood cavity	420	2
Left ventricle blood cavity	500	3
Right atrium blood cavity	550	4
Right ventricle blood cavity	600	5
Ascending aorta	820	6
Pulmonary artery	850	7

Table B.1: Whole heart substructures of images of the MM-WHS dataset with the corresponding original labels and newly assigned labels.