Deep MR to CT Synthesis for PET/MR Attenuation Correction

Kerstin Kläser

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Medical Physics and Engineering

University College London

Monday 9th August, 2021

I, Kerstin Kläser, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

"Sometimes it is the people no one can imagine anything of who do the things no one can imagine." – Alan Turing

Abstract

Positron Emission Tomography - Magnetic Resonance (PET/MR) imaging combines the functional information from PET with the flexibility of MR imaging. It is essential, however, to correct for photon attenuation when reconstructing PETs, which is challenging for PET/MR as neither modality directly image tissue attenuation properties. Classical MR-based computed tomography (CT) synthesis methods, such as multi-atlas propagation, have been the method of choice for PET attenuation correction (AC), however, these methods are slow and suffer from the poor ability to handle anatomical abnormalities. To overcome this limitation, this thesis explores the rising field of artificial intelligence in order to develop novel methods for PET/MR AC.

Deep learning-based synthesis methods such as the standard U-Net architecture are not very stable, accurate, and robust to small variations in image appearance. Thus, the first proposed MR to CT synthesis method deploys a boosting strategy, where multiple weak predictors build a strong predictor providing a significant improvement in CT and PET reconstruction accuracy.

Standard deep learning-based methods as well as more advanced methods like the first proposed method show issues in the presence of very complex imaging environments and large images such as whole-body images. The second proposed method learns the image context between whole-body MRs and CTs through multiple resolutions while simultaneously modelling uncertainty.

Lastly, as the purpose of synthesizing a CT is to better reconstruct PET data, the use of CT-based loss functions is questioned within this thesis. Such losses fail to recognize the main objective of MR-based AC, which is to generate a synthetic CT that, when used for PET AC, makes the reconstructed PET as close as possible to the gold standard PET. The third proposed method introduces a novel PET-based loss that minimizes CT residuals with respect to the PET reconstruction.

Acknowledgements

This work was sponsored by Siemens Healthineers, the UCL Impact Scheme, and the EPSRC-funded UCL Centre for Doctoral Training in Medical Imaging. I am thankful for their financial support and for creating an interdisciplinary environment which I greatly benefited from.

In addition, I would like to thank a number of people personally that contributed greatly to this thesis and my journey of becoming a PhD. First of all, I would like to thank my primary supervisor Sebastien Ourselin, for trusting me since day one of this journey and always having my back. Thanks for building a research group that I see as one big office family. I also thank my second supervisor, Jorge Cardoso, who walked this road with me every step of the way, believing in me from the beginning and being an excellent mentor. Thank you for encouraging me at the lows of this journey and celebrating the highs together with me.

I also want to thank each member of the AMIGO team who made coming to the office special every day. This thesis would have not been possible without the many valuable discussions, daily good morning hugs and the endless snack supply before a deadline. You are great colleagues, but even better friends.

A special thanks goes to Marta, Nooshin, Alessia, Minitom, Pedro, Zach, Graham, Jose and Loic. Words cannot describe the gratitude I feel for the support and love I have received from you. No matter the time you have walked this road by my side, you made an impact. You were my biggest allies and you will always have a place in my heart.

Furthermore, I want to thank Sarah and Rahima. You are the brightest rays of sunshine in the office family. You were always there for me, to listen, to give grownup advice and to put a smile on my face when things got stressful.

Thank you to my friends from the gym, Kate, Sian, Pennie, Dave and Gemma. You were my support bubble outside of the office. You often helped me to forget about work

and enjoy the moment.

Thank you, Alex and Meg, for being the most wonderful housemates one could imagine. You made Wightman Road a home and lockdown without you by my side would have been unbearable.

I want to finish this section by expressing my infinite gratitude to the beautiful humans who have supported me from all over the world and have shown me that friendship and love do not know distance.

Thank you to my Klassenfahrt Crew, Nana and Yannick, Eli and Ulrich, Lisa and Simon, Jill and Chris, Thomas and Jan. No matter how far we live apart, you will always be my home. The bond we built since we were little is a once-in-a-lifetime kind of friendship and I cannot wait to share more adventures with you. You are simply the best.

Thank you, Claudia and Daniel, for being like older siblings to me. You are the kindest and most wonderful humans that I have ever met. You have accompanied me on my way from highschool, to university to becoming a PhD. Your support and most genuine friendship means the world to me.

Thank you to Mona and Inga, my dearest friends and best travel companions. You are the most understanding and supportive friends one can ask for. I am forever grateful that I have met you in Lübeck. With you by my side everything seems not just doable, but easy and joyful. I will always treasure your loving souls and our close friendship.

The most special thank you goes to Jorge. You have become my safe harbor since the moment you stepped foot into my life. Your unconditional love and support has given me the stamina to finish this marathon. You are the minion that worked so hard in the background to make this thesis possible. I would not be here without you.

Finally, I want to thank my family from the bottom of my heart. Thank you Mama, for making me the person I am today. You have taught me to be a strong and independent woman and have shown me that I can achieve anything if I work hard for it. Thank you Paps, for always believing in me, being the proudest father one could imagine and for inheriting me your charme. Thank you Martin, for being the most protective big brother and teaching me the Rule of Three at an age of six. Thank you to Sanela, your sweet soul made my brother an even better brother.

Without the support and love of my beloved family and friends, I would not be where I am today, this is also your achievement.

Journal Papers

K. Kläser, P. Borges, R. Shaw, M. Ranzini, M. Modat, D. Atkinson, K. Thielemans, B. Hutton, V. Goh, G. Cook, M. J. Cardoso, and S. Ourselin. A multi-channel uncertainty-aware multi-resolution network for mr to ct synthesis. *Applied Sciences*, 11(4):1667, 2021a.

K. Kläser, T. Varsavsky, P. Markiewicz, T. Vercauteren, A. Hammers, D. Atkinson, K. Thielemans, B. Hutton, M. J. Cardoso, and S. Ourselin. Imitation learning for improved 3d pet/mr attenuation correction. *Medical Image Analysis*, 71:102079, 2021b.

J. Lillington, L. Brusaferri, **K. Kläser**, K. Shmueli, R. Neji, B. F Hutton, F. Fraioli, S. Arridge, M. J. Cardoso, S. Ourselin, K. Thielemans, and D. Atkinson. Pet/mri attenuation estimation in the lung: A review of past, present, and potential techniques. *Medical physics*, 47(2):790–811, 2020.

Conference Papers

K. Kläser, P. Borges, R. Shaw, M. Ranzini, M. Modat, D. Atkinson, K. Thielemans, B. Hutton, V. Goh, G. Cook, M. J. Cardoso, and S. Ourselin. Uncertainty-aware multi-resolution whole-body mr to ct synthesis. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 110–119. Springer, 2020.

M. B. M. Ranzini, I. Groothuis, **K. Kläser**, M. J. Cardoso, J. Henckel, S. Ourselin, A. Hart, and M. Modat. Combining multimodal information for metal artefact reduction: An unsupervised deep learning framework. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 600–604. IEEE, 2020.

K. Kläser, T. Varsavsky, P. Markiewicz, T. Vercauteren, D. Atkinson, K. Thielemans, B. Hutton, M. J. Cardoso, and S. Ourselin. Improved mr to ct synthesis for pet/mr attenuation correction using imitation learning. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 13–21. Springer, 2019.

K. Kläser, P. Markiewicz, M. Ranzini, W. Li, M. Modat, B. F. Hutton, D. Atkinson, K. Thielemans, M. J. Cardoso, and S. Ourselin. Deep boosted regression for mr to ct synthesis. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 61–70. Springer, 2018.

C. J. Scott, J. Jiao, M. J. Cardoso, **K. Kläser**, A. Melbourne, P. J. Markiewicz, J. M. Schott, B. F. Hutton, and S. Ourselin. Short acquisition time pet/mr pharmacokinetic modelling using cnns. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 48–56. Springer, 2018.

Impact Statement

One of the main challenges that PET/MR faces is the ability to directly correct for tissue attenuation, which is essential in order to reconstruct quantitative PET images necessary in clinical practice, e.g., to monitor disease progression.

In the scope of this thesis, multiple deep learning based methodologies have been developed in order to provide a fast and reliable means for PET/MR attenuation correction. At the beginning of this thesis, only a few groups in the medical imaging field had attempted to utilize convolutional neural networks in order to generate pseudo CT images from MR input images that can be linearly rescaled and used as attenuation maps. The methodologies proposed in this thesis reach from improving attenuation correction in the brain over the mainly uncharted territory of whole-body pseudo CT synthesis to the proof-of-concept of a novel loss metric that is the first of its kind.

All proposed convolutional neural networks have been thoroughly evaluated on hold-out datasets. Both quantitative and qualitative pseudo CT synthesis performance was improved by incorporating techniques known from classical machine learning, learning from multiple resolutions and even by having the networks imitate the PET reconstruction process. The proposed Deep Boosted Regression network achieves state-of-the-art results that almost reach the theoretical limit of two CT scans that were acquired consecutively. As a means of safety and in order to know the network's prediction confidence, the proposed multi-resolution network for whole-body MR to CT synthesis was extended to be able to model two kinds of uncertainty, thus accounting for intrinsic data noise and model uncertainty. The proposed Imitation Learning method was further evaluated on an independently acquired dataset exploiting the generalizability and the extrapolation properties of the novel metric loss that in itself imitates the PET reconstruction process.

With the increasing interest in fast and reliable attenuation correction methods that can be incorporated into clinical settings, the proposed methods could be integrated into com-

mercial software. This would be of particular interest for the application of whole-body PET/MR attenuation correction as the use of whole-body PET/MR imaging in current clinical settings is almost non-existent due to insufficient attenuation correction methods. Making models uncertainty-aware can further provide a measure of safety indicating whether the results are to be trusted and thus giving clinicians the possibility to recourse to other methods implemented on the scanner.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Clinical background

Medical imaging is an essential part of clinical analysis and medical intervention and aims to improve patient care by preventing, detecting, diagnosing and monitoring medical conditions. Some of the most popular imaging modalities used in clinical routines include magnetic resonance (MR) imaging, computed tomography (CT) and positron emission tomography (PET). While MR and CT are able to give an insight into the anatomy of a patient, PET provides information about the metabolic functionality within the patient's body. In recent years, researchers have focused on combining the advantages of these imaging techniques. Most recently, researchers were able to combine MR with PET surpassing multiple engineering challenges. Combining the excellent soft tissue contrast of MR images and the functional image information from PET yield hopes to significantly improve clinical practice and offer a new range of clinical applications. There are, however, some limitations with this technique that need to be overcome. One way to tackle these limitations has focused on techniques from machine learning, in particular deep learning. Deep learning has proven to successfully solve complex problems in the field of computer science/vision. The roots of deep learning are inspired by the way neurons in the human brain work. To have a better understanding of each imaging modality and the theory behind deep learning algorithms, the following chapter explains the physics behind each imaging modality and the basic ideas and algorithms used in deep learning.

1.2 Basic physics and technology

To better understand the context of this work, this section explains the physical concepts behind the imaging modalities used throughout the thesis. Furthermore, an introduction is

given into the deep learning concepts which are essential building blocks of this work.

1.2.1 MR concepts

MR imaging is an imaging technique that makes use of the magnetic characteristics of protons, usually of the least complex element, hydrogen (H), in order to generate an image. The single proton in the nucleus of the hydrogen molecule does not sit statically in the centre of the atom, but rotates on its axis, creating the so called spin, which is associated with all protons within the body. In their entirety, they generate a small magnetic field, also known as magnetic moment. The spin magnet's orientation is random and therefore zero in total under normal circumstances. However, exposing the protons to an external magnetic field ($B0$) causes all spins to line up either in the same or opposite direction of the magnetic field depending on their energy state. The spins now precess around the $B0$ axis, commonly referred to as the z-axis. There is no magnetization in the transverse x-y-plane as the spins cancel each other out within this plane but they sum up along the z-axis resulting in a net magnetization $M0$ which is proportional to the proton density (PD).

The basic idea behind magnetic resonance is to perturb the equilibrium of the spins. Disequilibrating the spins makes it possible to detect a change in magnetization. A high-frequency (HF) pulse is transmitted into the vicinity of the protons in order to throw the protons out of their equilibrium state. The strength and duration of the HF pulse determine the angle at which spins and therefore net magnetization are flipped. Now all spins are exactly in phase and a net magnetization in the transverse direction Mxy can be detected whereas the magnetization along the z-axis (Mz) vanishes. When turning off the HF pulse the spins return to their resting alignment through various relaxation processes. The time it takes the spins to regrow the longitudinal magnetization Mz to the original magnitude $M0$ is called T1, also known as longitudinal or spin-lattice relaxation time. The transverse relaxation time T2 describes the time it takes until the transverse magnetization (Mxy) has decayed. This phenomenon happens because the spins lose the phase coherence as some precess slightly faster than others due to natural cross-talk between neighboring spins. It is therefore also referred to as spin-spin relaxation time. Inhomogeneities in the main magnetic field $B0$ cause Mxy to decrease faster. This "observed" or "effective" T2 time constant is called T2*. MR sensors are able to detect the transverse magnetization as an electrical current. from which an image can be reconstructed.

However, it is not possible to distinguish between different tissue structures yet due to

missing spatial allocation. Therefore, MR makes use of additional gradient fields that are generated by a pair of gradient coils in x-, y- and z-direction with the same current strength, but opposite polarity, making it possible to localize each individual voxel. The detected signals are then saved line by line in a raw data matrix, the so called k-space. The final MR image is reconstructed from the k-space by means of a 2D Fourier transform. In general, low spatial frequencies are responsible for the MR contrast and high spatial frequencies determine small structures like tissue boundaries.

Two important pulse sequence parameters that are chosen by the operator are repetition time (TR) and echo time (TE). These parameters determine the MR image contrast resulting in T1-, T2- or PD-weighted images. TR describes the time between two HF pulses and TE is the time between applying the HF pulse and the peak of the signal induced in the coil. Choosing a short TR and TE results in a T1-weighted image as T2 effects largely disappear, whereas a T2-weighted image is generated when TR and TE are set to a longer time as T1 effects disappear. PD-weighted images are a result of a long TR and a short TE, which minimizes both T1 and T2 effects. Figure 1.1 shows the correlation between TR/TE and resulting contrast.

In PD images, proton-rich tissues (e.g., fat, fluids, grey matter) generate a high signal and appear bright in the resulting image. These kind of images are very useful for brain

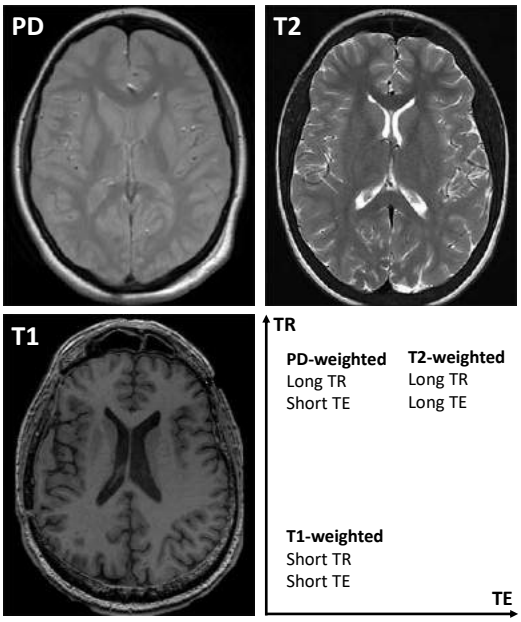

Figure 1.1: Three different MR contrasts: T1, T2 and PD. Repetition time (TR) and echo time (TE) determine resulting image contrast.

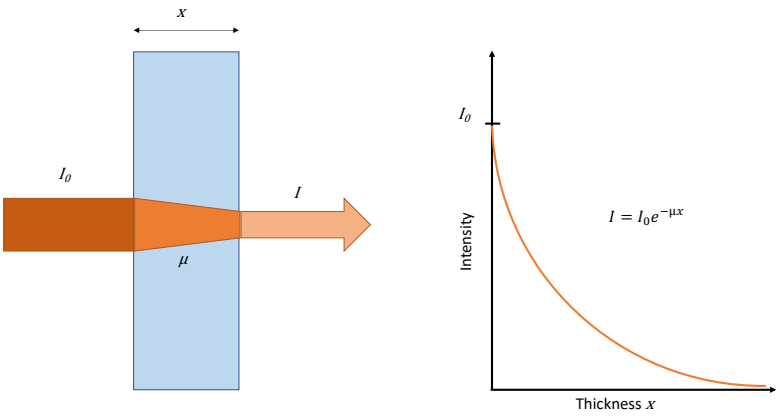

Figure 1.2: X-ray attenuation: the intensity I_0 of an initial X-ray beam is partly absorbed by an object resulting in an attenuated X-ray beam with intensity I. The intensity reduction follows Beer's Law, which is an exponential function of X-ray energy (I_0), path length (x), and material specific attenuation coefficient (μ).

imaging owing to the great contrast between grey and white matter. T1-weighted images on the other hand are characterized by the dark appearance of fluid filled spaces in the body (e.g., Cerebrospinal fluid in the brain, free fluid in the abdomen, fluid in the gall) and are very useful for brachial and lumbar plexus imaging, brain imaging and extremity imaging. Lastly, T2-weighted images are often used for abdominal imaging, pelvic imaging and chest imaging as they brightly depict pathological processes.

1.2.2 CT concepts

CT, like X-ray imaging, is based on the fundamental principle that the X-ray absorption of different tissues is variable. In CT the transmitted intensity of a beam of radiation that is finely collimated is measured making cross-sectional imaging possible. When an X-ray beam hits an object, its energy (I_0) is partly absorbed by the object. Thus, the beam that hits the detector has a decreased intensity (I). This decrease follows the Lambert-Beer Law, which describes the intensity reduction as a function of X-ray energy, path length (x) and material specific attenuation coefficient (μ). Figure 1.2 illustrates this concept.

The image itself is reconstructed by solving the inverse problem to compute the attenuation coefficients along different projection lines. A CT image consists of an array of pixels representing the mean attenuation within each pixel that is expressed as a *Hounsfield Unit* (HU) (Air = -1000, Fat = -60 to -120, Water = 0, Compact bone = +1000). Hounsfield Units are defined as

$$HU_i = 1000 * \frac{\mu_i - \mu_{water}}{\mu_{water}}, \tag{1.1}$$

where μ_i describes the object's attenuation coefficient in the i-th voxel and μ_{water} the linear attenuation coefficient of water. Since the tissue density determines the degree to which the X-rays are attenuated, the reconstructed image is bright in tissues with a high attenuation coefficient and dark in those that absorb with low attenuation coefficients.

1.2.3 PET concepts

PET is a nuclear imaging technique that relies on the decay characteristics of radionuclides which decay after emitting a positron (β^+). Within the decay process of a radioactive atom, a positron that was ejected from the nucleus covers a short distance within the tissue until it slows down and interacts with an electron. This collision annihilates both electron and positron, resulting in a pair of high-energy photons (511 keV) with opposite directions (almost 180° apart). These photons have a high probability of escaping the body. The distance that a positron travels before it collides with an electron depends on its kinetic energy; positrons with high kinetic energy will travel further than the ones with low kinetic energy. The annihilation process is shown in Figure 1.3.

In order to reconstruct a PET image, the two annihilation photons with opposed directions are detected by two opposed detectors. When both detectors register a signal within a given time window, a coincidence event is saved in a data matrix. In time of flight (TOF) PET systems, an additional parameter is saved in the data matrix: the time difference between the two detected annihilation photons. This additional information makes it possible to reduce the annihilation origin to a limited range. The resulting data matrix does not represent the radiotracer distribution directly, which is why a reconstruction process is essential to determine the radioactive tracer distribution. PET reconstruction can broadly be split into analytical (e.g. filtered backprojection) and iterative reconstruction algorithms (e.g. expectation maximization (EM)). To accurately quantify the radionuclide uptake it is essential to correct for multiple factors including tissue attenuation, which represents a major image degrading factor. Annihilation photons will be attenuated most when travelling through bone and least when travelling through air. Without attenuation correction, false coincidence events are likely to be saved in the data matrix resulting in inaccurate PET reconstruction.

In general, PET is used to examine physiological and biochemical functions of the body. Radionuclides are attached to substances that take part in metabolic functions. Due to

the fact that radioactive isotopes have the same chemical properties as their non-radioactive counterparts, those labeled substances are metabolized identically.

1.2.4 Combined imaging modalities

PET images provide information about radionuclide uptake distribution within the body, however, PET is not able to provide any anatomical information, which can lead to wrong interpretations of the precise location of areas with abnormal radionuclide uptake. In additon, one of the most commonly used tracers, Fluorodeoxyglucose (^{18}F-FDG), is metabolized not only in unhealthy tissue but in multiple healthy organs such as the brain, lungs, heart, intestines and liver. To improve interpretability of PET images it is therefore desirable to combine PET images with an anatomical imaging modality such as CT or MR. Figure 1.4 shows example images for PET, CT and MR images and their combinations.

1.2.4.1 PET/CT

In 2001, the first commercial PET/CT scanner was installed facilitating dual-modality imaging (Beyer et al. 2000). Both PET and CT scanners are combined in a single gantry that

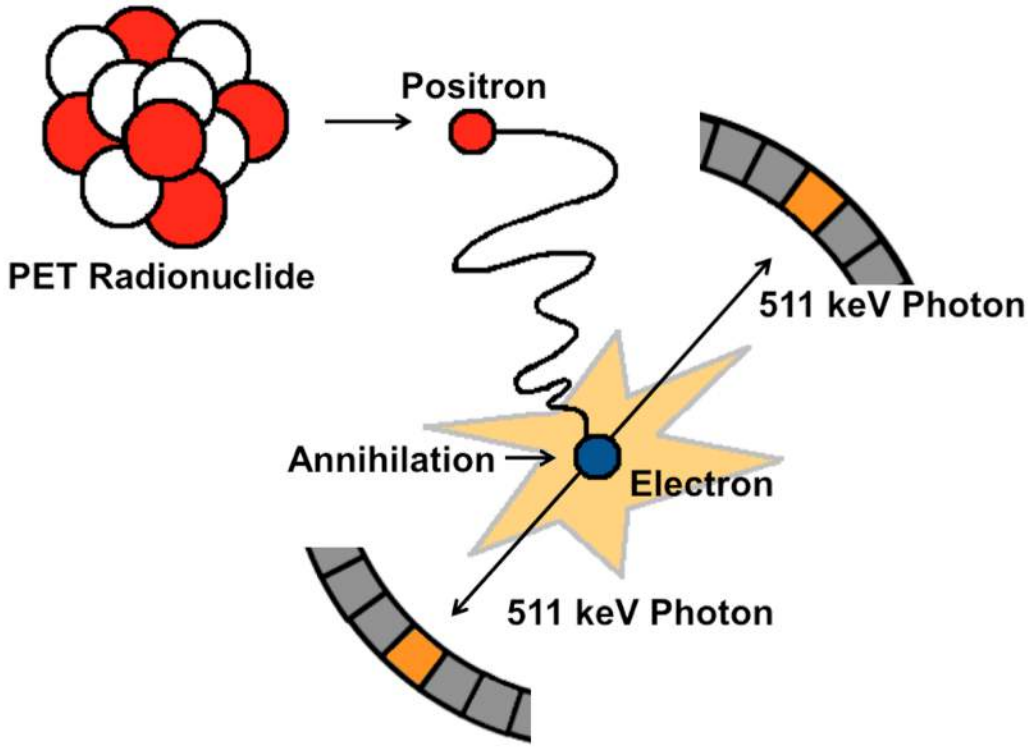

Figure 1.3: Process of PET annihilation. The radionuclide decays and a positron travels for a short distance until it slows down and interacts with an electron resulting in a pair of high-energy photons in almost exactly opposite directions.

surrounds the patient bed. Images are acquired sequentially from both imaging devices in a single session resulting in co-registered PET and CT images without the need to move the patient. PET/CT scanners allow a combined anatomical and functional examination of the patient in one imaging session and help radiologists interpret the results more precisely. Clinical practice has become much easier for patients and doctors with the emergence of PET/CT systems and has therefore exclusively been used for PET acquisitons since 2006 (Townsend 2008).

1.2.4.2 PET/MR imaging

In 2010 the first whole-body PET/MR scanner was commercially available. In fact, two different systems were brought to the market by manufacturers Philips (Philips Medical Systems, Best, The Netherlands) and Siemens (Siemens Healthcare, Erlangen, Germany) in the same year. The two competing systems differ mainly in the image acquisition technique. The Ingenuity TF PET/MR from Philips acquires PET and MR images sequentially. The system consists of an MR scanner and a separate PET scanner in the same room with a rotating bed in between so that the patient does not need to be transferred between scans. On the contrary, there is the mMR Biograph from Siemens which is a fully integrated system that allows simultaneous PET and MR acquisition. Building such a simultaneous system is technically challenging due to multiple factors. On one hand, there are spatial constraints due to the limited bore size of the MR scanner that make it difficult to integrate a rotating transmission source within an MR gantry. On the other hand the strong magnetic field can interfere with the detection of the PET signal due to the high sensitivity of the photomultiplier tubes (PMTs) of the PET detectors to magnetic fields. However, simultaneous systems provide the advantage of the ability to make simultaneous, exactly aligned, acquisitions (Daftary 2010).

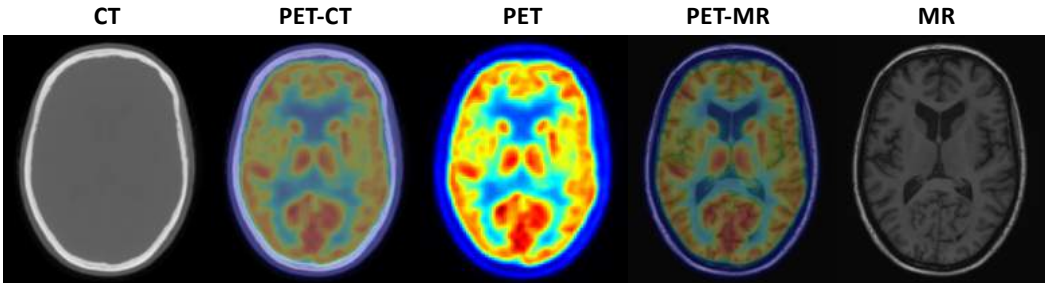

Figure 1.4: PET, CT and MR images and their corresponding fusions.

1.2.5 Attenuation correction PET

In PET, attenuation from the object inside the field of view (FOV) is one of the major image degrading factors. In order to avoid image distortions and artefacts, attenuation correction is an important component of PET image reconstruction. Without attenuation correction it is not possible to accurately quantify regional tracer uptake for routine clinical studies and for quantitative dynamic studies. An extensively used attenuation correction method in standalone PET scanners is based on transmission measurements. However, there are several difficulties to implement a rotating transmission source within a PET/MR scanner due to the restricted bore size and the strong magnetic field. With the emergence of PET/CT scanners, attenuation correction became straightforward since a map of linear attenuation coefficients at 511 keV in the object can be rapidly derived from a CT scan using piece-wise linear calibration curves. However, MR image intensity values are related to proton density and do not provide information about X-ray attenuation. This becomes obvious with respect to bone and air, which yield a similar MR signal for many sequences, but have very different attenuation coefficients. Therefore, the development of alternative attenuation correction methods has become a main field in PET/MR research. Methods for PET/MR attenuation correction can be categorized into five classes based on the techniques applied to create the attenuation map, also referred to as μ-map: transmission-, emission-, segmentation-, atlas- and learning-based approaches. A thorough review of available attenuation correction methods for PET/MR imaging is covered in chapter 2.

1.3 Deep learning concepts

Deep learning is an area within machine learning research which has gained a substantial amount of attention in recent years. In deep learning, several levels of representation and abstraction are learned in order to connect data (e.g., images, sound, text). A deep learning algorithm does not know the answer to a specific task straight away. In order to train a deep learning network, the algorithm instead continuously analyzes training data and adapts its approach depending on its performance.

Artificial neural networks are one main element within the field of deep learning. Neural networks, as the name may give away, are computational models inspired by the way the cerebral cortex processes information in the human brain. Mathematically, biological neurons are represented by computational units called perceptrons that can be connected to each other across several layers. On its way through the network, the data runs from an input

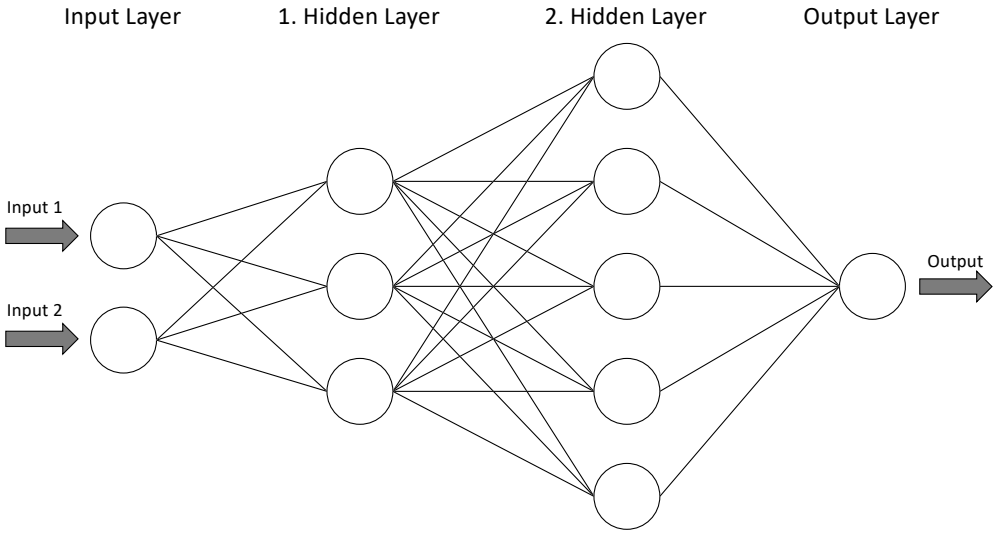

Figure 1.5: A simple feed-forward neural network with an input layer, two hidden layers and an output layer.

layer through a collection of *hidden layers* that consist of multiple perceptrons, or nodes, until it reaches the output layer. Such networks are also known as *feed-forward networks*, which means that information is always fed forward, never back. Figure 1.5 shows a simple example of such a neural network. If every node in a given layer passes its output to every node in the following layer, they are referred to as *fully connected layers*. In this case the output layer neurons do not have any direct connection to the input data and are only activated by the signal of the previous layer. The more layers of abstraction a neural network is built of, the deeper it is and the more computational resources are needed to generate the output. In Fig. 1.5, the input layer and the output layer are connected by two hidden layers that consist of three and five nodes respectively. Therefore this example neural network has a depth of two.

Each connection between nodes of adjacent layers has a weight θ_i associated with it. This weight determines to what extent the input to the node contributes to the output of the node. When training a neural network, the model tries to determine the set of weights that give the best output while keeping the network architecture unchanged. In Fig. 1.5 the weights are represented by the lines between different nodes. If there are multiple inputs to

a node, the output of a given node is commonly calculated by a weighted sum defined as $z = \sum_{i=1}^{n} \theta_i x_i$. The input data x_i is multiplied by the weight θ_i that is associated with each individual input in order to compute the output z. The final output y of the node, however, consists of two more components: a bias b and an activation function f, such that

$$y = f(\sum_{i=1}^{n}(\theta_i x_i) + b). \tag{1.2}$$

The bias can be interpreted as the intercept of a linear function. It allows the activation function to shift left or right so that it does not necessarily go through the origin. The activation function determines if a node activates, similar to a neuron "firing" in a biological context. If a neuron's input is relevant for the model's prediction, the node activates. The activation function can be considered as a "gate" between an input for a node and its output. The complexety of this function can range from a simple step function through linear functions to more complex non-linear functions. Figure 1.6 shows four examples for possible activation functions. Step functions have a binary output of 0 or 1 and can be switched on or off. The use of a binary step function as activation is often too simple and does not allow for multi-value outputs. This, for example, can be problematic when trying to classify an input into several output categories. Linear functions $f = cz$ allow for multiple outputs by generating an output signal that is proportional to the node's weighted sum. However, a linear activation function causes the output layer to be a linear function of the inputs and therefore is not able to grasp the complexity of non-linear problems. Non-linear activation functions, on the contrary, allow for the neural network to represent data in a more complex way. Commonly used non-linear activation functions include sigmoid and rectified linear unit (ReLU) functions. The sigmoid function is defined as

$$f = \frac{1}{1 + e^{-z}}. \tag{1.3}$$

It maps a real-valued input and squeezes it to a range between 0 and 1. When used by each node in a multi-layer neural network, the sigmoid activation function generates a new "representation" of the original data that cannot be represented by any linear combination of the input data. Unfortunately, a common problem that can be observed when training a network with sigmoid activation functions is the so called vanishing gradient problem (more about gradients in subsection 1.3.1.1). This phenomenon can occur for very high

or low values of z and results in a stalling network that does not learn any further. The ReLU activation function circumvents the vanishing gradient problem by taking a linear activation function and setting it to zero for $z <= 0$, such that complex relationships can still be modelled. The ReLu is defined as

$$f = max(0, z) \tag{1.4}$$

and is called a piecewise function due to the combination of linear and non-linear sub-functions. It results in an output z if z is positive and 0 otherwise. Glorot et al. (Glorot et al. 2011) have shown that the use of ReLU improves learning in networks with three or more hidden layers, which makes it one of the most popular choices among activation functions.

The idea behind deep neural networks is to simplify a complex task by dividing it into multiple simpler tasks. Each hidden layer by itself represents a different task, or function, but when concatenated they can describe the more complex task. The depth and complexity of a neural network architecture therefore determine the level of task complexity that can be solved. However, increasing network depth comes with the pitfall of more network weights that must be learned, which can quickly turn into a computational burden. When designing a neural network one tries to find the best compromise between depth and computational requirements.

1.3.1 Optimizing neural networks

Just like in classic machine learning techniques, there are multiple ways to train a network depending on the relation between input and output data. The simplest form of learning is called supervised learning and refers to the scenario where paired data is available. In a medical context this could be in the form of an electroencephalogram (EEG) signal and a

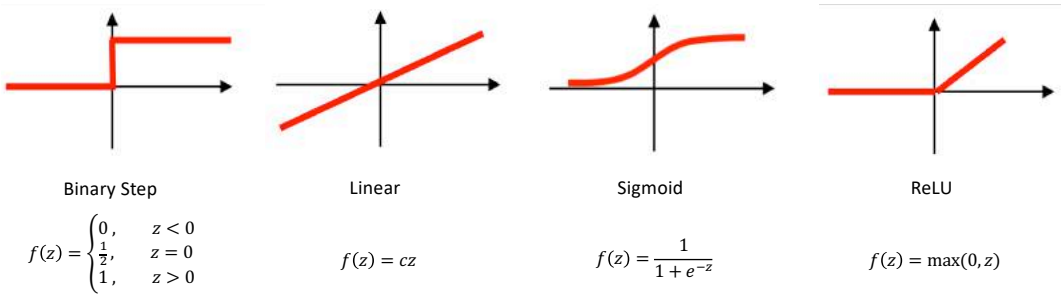

Figure 1.6: Activation functions for neurons. Left to right: binary step function, linear function, sigmoid function, rectified linear unit (ReLU).

corresponding diagnosis for training a classification network; MR images of the abdomen and a corresponding binary organ mask for segmentation purposes; or paired MR and CT images in order to perform an image regression task. Depending on the task, the neural network tries to predict either a class or set of classes (classification), a binary pixel value (segmentation) or a continuous pixel value (regression). No matter the task, the training procedure follows the same principle. When a network is trained from scratch, the weights are randomly initialized at the beginning of training and will most likely give an equally random output. The network's weights are then automatically updated during training in order to find a set of weights that give the best prediction whilst keeping the architecture unchanged. In order to evaluate its performance, neural networks make use of objective functions. These objective functions project real error values as a function of the weights in a multi-dimensional feature space. A typical choice for objective functions are loss functions that estimate the error between a prediction and the ground truth image. The goal is to minimize this error such that the output prediction becomes more accurate after adjusting the network's weights. Each change in the network's weights results in a loss function change and therefore represents a gradient. If the gradient is negative and points towards the steepest descent the loss function decreases as quickly as possible resulting in a network output that is closer to the ground truth than the previous prediction. Two popular optimization techniques that are commonly used in neural networks are stochastic gradient descent (SGD) and the Adam optimizer.

In SGD a small number of samples n are randomly selected from the data-generating distribution and grouped into so called mini batches. Instead of calculating the gradient g on the whole training sample distribution, which is computationally expensive and often unfeasible, SGD adjusts the network's weights according to the gradients of the mini batch. This iterative process is repeated until the value of the loss function $\mathcal{L}$ no longer decreases, meaning a minimum has been found. Algorithm 1 summarizes the SGD method.

Each mini batch includes n examples from the training data $x^{(1)},...,x^{(n)}$ with corresponding labels/ground truth $y^{(i)}$. The gradients of the loss function are computed at every iteration and the difference between the weights and the gradients determines the updated weights. The magnitude and direction of the weight update are computed by taking a step in the opposite direction of the loss gradient. In practice, the gradients are often too big and are therefore scaled by the parameter η, also referred to as step size or learning rate. The learn-

Algorithm 1: Stochastic Gradient Descent

η = Learning rate
θ = Weights (Randomly initialized)
$\mathcal{L}$ = Loss function
n = Mini batch size
while θ *has not converged* **do**
  $g \leftarrow \frac{1}{n}\nabla_\theta \sum_i \mathcal{L}(f(x^{(i)};\theta), y^{(i)})$ (Compute gradient)
  $\theta \leftarrow \theta - \eta g$ (Update weights)
end

ing rate is a critical hyperparameter in the training process and must be chosen carefully in order to avoid overfitting. Typical values for η are between 1.0 and 10^{-6}. Choosing a fixed learning rate can lead to the false belief that the model has already converged. Therefore, it is useful to gradually decrease the learning rate over time.

Another popular optimization algorithm has been introduced by Kingma and Ba (Kingma & Ba 2014) called Adam, which stems from adaptive moment estimation. It is different to SGD as it introduces momentum and an adaptive learning rate in order to find individual learning rates for each weight parameter. SGD does not consider any of the previous steps whereas Adam introduces an exponentially decaying average of past gradients. The use of momentum is desirable in order to avoid getting stuck in a local minimum. Algorithm 2 summarizes the Adam optimization method. The monotonous step size used in SGD is adapted such that it incorporates the momentum of prior steps, where parameters m and v are the estimates of the first (mean) and second (the uncentered variance) momentum of the gradients respectively. To estimate m and v, Adam makes use of the exponentially moving averages of the gradient and corrects for initial bias. After incoorporating the bias-corrected momentum estimates ($\hat{m}$, $\hat{v}$) the weight update rule changes to $\theta \leftarrow \theta - \eta \frac{\hat{m}}{\sqrt{\hat{v}}+\varepsilon}$, where ε is a small constant that prevents division by zero.

The authors of the original work recommended the following values for the parameters of Adam: $\eta = 10^{-3}$, $\varepsilon = 10^{-8}$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Adam is the optimizer that has been used in all experiments described in the following chapters. When training a neural network the optimization is run until a stopping criterium is met. Usually, this stopping criterium is either a pre-defined number of traning iterations or until convergence of the loss function, where convergence is typically defined as a sub 5% change in the loss value over a period of 5000 iterations.

Algorithm 2: Adam

η = Learning rate
θ = Weights
$\mathcal{L}$ = Loss function
n = Mini batch size
$\beta_1, \beta_2 \in [0;1)$ = Exponential momentum decay rates
ε = Constant to ensure numerical stability
$m \leftarrow 0$ (Initialise 1st momentum)
$v \leftarrow 0$ (Initialise 2nd momentum)
$t \leftarrow 0$ (Initialise time step)
while θ *has not converged* **do**
  $t \leftarrow t + 1$
  $g \leftarrow \frac{1}{n}\nabla_\theta \sum_i \mathcal{L}(f(x(i);\theta), y^{(i)})$ (Compute gradient)
  $m \leftarrow \beta_1 m + (1-\beta_1)g$ (Update 1st momentum estimate)
  $v \leftarrow \beta_2 v + (1-\beta_2)g^2$ (Update 2nd momentum estimate)
  $\hat{m} \leftarrow \frac{m}{(1-\beta_1)}$ (Correct bias in 1st momentum)
  $\hat{v} \leftarrow \frac{v}{(1-\beta_2)}$ (Correst bias in 2nd momentum)
  $\theta \leftarrow \theta - \eta \frac{\hat{m}}{\sqrt{\hat{v}}+\varepsilon}$ (Update weights)
end

1.3.1.1 Backpropagation

Another main component of training neural networks is an algorithm called *backpropagation*. It is used to evaluate the gradients used in the aforementioned optimization strategies. Backpropagation gained popularity in 1986 when David Rumelhart, Geoffrey Hinton, and Ronald Williams discovered that neural networks can be trained much faster using backpropagation than with earlier approaches (Rumelhart et al. 1986). Like with many other algorithms, the methodological idea of backpropagation is encapsulated in its name. The gradients of the loss function are propagated back through the network using the chain rule of calculus. This means that the gradients are the change in loss with respect to the network weights, evaluated for the current input. The algorithm aims to compute the partial derivatives of the loss function with respect to any weight θ and any bias b. Mathematically speaking, the derivative of the loss function at a distinct point can be described as the rate at which the loss function is changing its value at this particular point. Neural networks often consist of many layers with associated weights, and therefore gradients, which is why the derivative of the loss function must be decomposed in order to determine the rate of change in every node of the network. Backpropagation is performed until each individual gradient with respect to weight and bias in the first layer is known. Once all gradients are known, they can be updated following the optimization strategies described in 1.3.1.

1.3.2 Convolutional neural networks

To this point, all neural networks described above are fully connected neural networks, meaning every node in each layer is connected to every neuron in the adjacent layer. The fully connected nature of such networks can quickly become a computational burden when working with large data. A typical three channel colour image of size $[256, 256, 3]$ would result in $[256 \times 256 \times 3] = 196608$ weights per node. Considering that fully connected neural networks are built out of multiple layers containing many neurons, the number of weights can easily become unfeasible to handle. In order to address this issue, convolutional neural networks (CNNs) were introduced that offer a more efficient way of dealing with large data (LeCun et al. 1989). Such networks make use of convolutional layers instead of nodes. Convolutional layers perform an operation called "convolution" similar to applying a filter to an array. Mathematically speaking, a convolution applied to a two-dimensional image I can be written as

$$S(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n) \tag{1.5}$$

where $K(m,n)$ is a two-dimensional kernel and $S(i,j)$ the resulting *feature map*. The idea can be visualized as sliding the kernel along the image I. An element-wise multiplication between the filter-sized patch of the input and the filter is performed and the multiplications are summed up into one scalar. This operation is repeated for each image position (i,j). The filter-sized patch of the input image is also called *receptive field*. Figure 1.7 illustrates an example of such a convolutional operation.

The values within the kernel are not fixed and can be adjusted during training similar to the weights in a fully-connected neural network. A CNN can be build out of many convolutional layers containing an arbitrary number of filters that each result in a different feature map. CNNs follow a hirarchical structure, which means that the abstraction of learned features increases for layers that are deeper in the network. The first layers learn simple features such as edges or corners, layers further down the stream learn abstract feature maps in which the human eye struggles to grasp the logical context. Convolutions are often followed by *pooling layers* that serve the purpose of decreasing the number of parameters while increasing the receptive field. Pooling can also be understood as a means of downsampling. Figure 1.7 bottom shows an example of a *max-pooling* operation. A kernel of size 2×2 slides over the image with *stride* 1. The stride indicates the number of

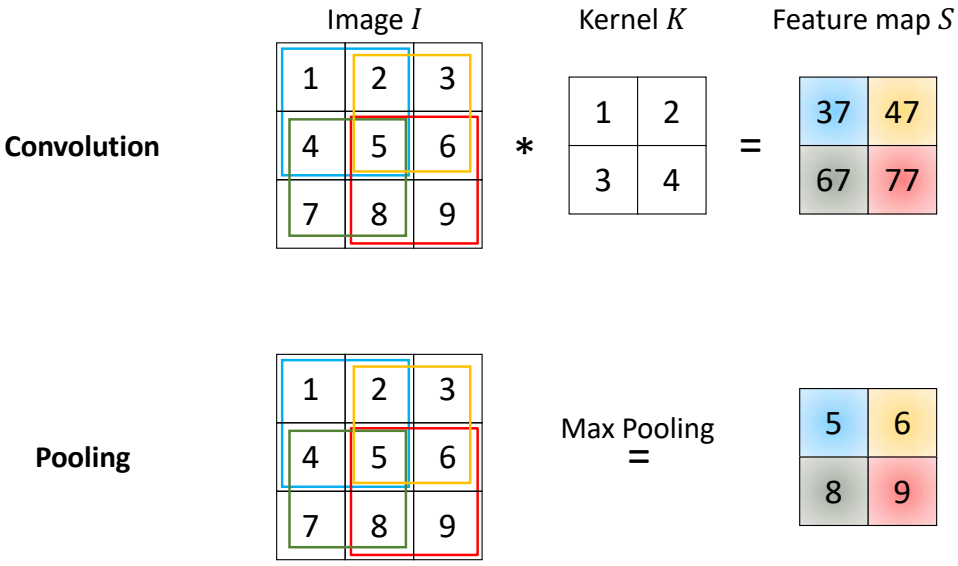

Figure 1.7: Top: 2D convolution. An element-wise multiplication between a filter-sized patch of the input image I and a kernel K is performed and the multiplications are summed up into one scalar resulting in a feature map S. Bottom: 2D max-pooling. Only the largest output within a rectangular neighborhood is kept in the feature map.

pixels that the kernel has moved after each feature map computation. In each feature map computation, represented by the coloured squares, the max function is applied to the values within the kernel and only the maximum value is kept resulting in feature maps that emphasize features like edges. Another popular pooling operation is *average pooling*, where the average function is applied within the kernel resulting in feature maps that smooth features like edges. It is important to note that the learning process of CNNs follows the same rules as fully-connected neural networks. Thus, the optimization strategies including SGD and Adam, as well as the backpropagation algorithm still apply for CNNs.

1.4 Thesis contribution

The ability to simultaneously acquire PET and MR images has the potential to pioneer new research and clinical applications especially in cases where information provided by a PET/CT acquisition is not sufficient. However, until today, PET/MR scanners have not yet reached the same recognition as PET/CT scanners predominantly due to the inferior performance of photon attenuation correction methods. A major problem that attenuation correction methods on PET/MR systems face is the lack of accuracy when reconstructing bone. This problem arises because air and bone have similar attenuation coefficients in most standard MR sequences, however, in reality, the two materials attenuate photons with op-

posite attenuation coefficients. Many groups have tried to solve this problem, especially in the brain where accurate quantification is especially important. For many years, multi-atlas propagation methods like Burgos et al. (Burgos et al. 2013) dominated the field due to their excellent performance and robustness. With the rising popularity of artificial intelligence in the field of computer sciences, new doors have opened for attenuation correction methods in medical imaging. Therefore, the first objective of this thesis is to develop a novel method for PET/MR attenuation correction in the brain that is based on deep learning.

Attenuation correction methods often focus on one particular body region only, for the most part, the brain. However, it is important to optimize PET/MR systems such that they are able to image any part of the body. This is of particular interest when acquiring a whole-body PET/MR image to detect metastases. However, whole-body images are large and prove to be problematic in deep learning based methods due to a limited GPU memory budget. Networks only see a limited part of the image and therefore struggle to capture the contextual information. An additional problem arises when creating a co-registered database because the patient's position differs in MR and CT scanners such that registration algorithms struggle to cope with the alignment of the images. The second objective of this project is therefore the development of a novel CT synthesis network for whole-body applications that further incorporates uncertainty estimations as a means of safety.

The most common way to optimize MR to CT deep learning algorithms is to minimize the error between the synthesized pseudo CT and the corresponding ground truth CT image, equivalent to minimizing the $\mathcal{L}_2$-loss. This objective is often justified by the fact that in current clinical practice the gold standard for PET/MR attenuation correction is an additional CT acquisition that can be linearly rescaled to an attenuation map used in PET reconstruction. However, $\mathcal{L}_2$-losses do not recognize that the main aim of CT synthesis, when used for PET/MR attenuation correction, is to generate a synthetic CT that, when used as attenuation map for PET reconstruction, makes the reconstructed PET as close as possible to the gold standard PET reconstructed with the true CT. Thus, the third objective of this thesis is to develop a novel deep learning method for MR to CT synthesis that directly minimizes the PET residuals when the pseudo CT is used for PET reconstruction.

The contributions include:

1. A novel CNN, namely Deep Boosted Regression (DBR), for CT synthesis of the head. The method was inspired by the success of deep neural networks in the field of

segmentation such as U-Net (Çiçek et al. 2016) and HighRes3DNet (Li et al. 2017) as well as *Boosting* known from classic machine learning (Schapire 1990). CNNs mimic the structure of the human visual cortex and are able to pick up patterns in the input image through a number of convolutions in order to make a prediction. Boosting combines several weak learners (here each represented by a CNN) that in their entirety build a strong learner. Instead of predicting labels like in segmentation, continuous CT intensities are predicted by the network resulting in what will be referred to as a pseudo CT (pCT).

2. In order to find the best possible learning conditions for the proposed network and make it more flexible, the method is tested on multiple MR contrasts. The method is further extended to be able to take multiple contrasts as input.

3. A second CNN is proposed particularly designed for whole-body CT synthesis. The network operates at different levels of resolution in order to capture high-level and low-level features. Additionally, the proposed MultiRes$_{unc}$ network models two kinds of uncertainty. Including uncertainty in the network acts as a measure of safety and to account for intrinsic noise and misalignment in the data.

4. Finally, a third CNN architecture is introduced following an Imitation Learning strategy so that the CT synthesis process directly includes information about the PET error when the pseudo CT is used as attenuation map for PET reconstruction.

5. The end-to-end optimization framework is evaluated on an independantly acquired head dataset to investigate the robustnes of the proposed method. Furthermore the PET reconstruction accuracy is assessed for ^{18}F-FDG PET images by comparing the PET image reconstructed with the attenuation map derived from the synthesized pseudo CT against the reference PET that was reconstructed with the ground truth CT derived μ-map.

1.5 Thesis organisation

Research in the field of PET attenuation correction can be categorized in five classes: transmission-, emission-, segmentation-, atlas- and learning-based approaches. The following chapter thoroughly reviews the main methods for each category. Chapter 3 addresses the first steps of tackling the MR to CT image translation task in a deep learning manner

ranging from pseudo CT synthesis as a classification task to a novel residual learning CNN. Chapter 4 shows how pseudo CT synthesis results change depending on the MR sequence that is used as input (T1-, T2- and T1- & T2-weighted MR images). Chapter 5 shows how the proposed methods that were developed for brain application perform on whole-body images and presents a novel method that is optimized to deal with the large scale prevalent in whole-body images. In chapter 6, an end-to-end optimization approach is presented that questions the use of the traditional $\mathcal{L}_2$-loss when synthesizing pseudo CTs for the purpose of PET/MR attenuation correction, including an evaluation of the proposed neural network's performance on an independently acquired brain dataset. Finally, chapter 7 concludes this thesis and discusses potential future research directions.

Chapter 2

Attenuation correction for PET/MR scanners

In the early years of PET imaging, attenuation correction was performed using a rotating transmission source. However, it is technically difficult to integrate a rotating transmission source within an MR gantry due to the limited space. When PET/CT scanners were developed, the attenuation correction method of choice was to derive the attenuation coefficients from the CT scan using piece-wise linear calibration curves (Burger et al. 2002). However, MR image intensities are proton density-related and do not provide information about X-ray attenuation, which is why alternative attenuation correction methods must be developed to ensure accurate image quantification in PET/MR imaging. Methods for PET/MR attenuation correction can be categorized into five classes based on the techniques applied to create μ-maps: transmission-, emission-, segmentation-, atlas- and learning-based approaches. Within the next chapter, previous work on attenuation correction methods for PET/MR imaging will be described including an overview of their advantages and disadvantages.

2.1 Transmission-based attenuation correction

The first class derives information about the attenuation within the object using an external transmission source (usually a positron emitter) and directly results in linear attenuation coefficients at 511 keV. Research within this field mainly focuses on the design of new transmission sources.

In 1995, Meikle et al. (Meikle et al. 1995) introduced a method to measure and correct for attenuation in whole-body PET using simultaneous emission and transmission. However, due to the limited bore size and the strong magnetic field of PET/MR scanners the

introduction of a rotating 68Ge/Ga rod source in the scanner is difficult. Therefore, it is essential to develop new transmission sources that can be integrated within the small bore of the scanner. In 2012, Mollet et al. (Mollet et al. 2012) proposed the use of an annulus transmission source. Emission and transmission data are acquired simultaneously and can be separated by using a time-of-flight (TOF) classification approach. In TOF PET systems, the time difference between two detected annihilation photons is measured, which allows the annihilation origin to be reduced to a limited range. This results in a decreased spatial uncertainty and an increased signal-to-noise ratio. In 2014, Mollet et al. (Mollet et al. 2014) showed in a study including five human PET/MR and CT datasets that sufficient statistics can be obtained with an annulus-shaped transmission source to derive attenuation maps.

Another method to estimate an attenuation map was proposed in 2014 by Kawaguchi et al. (Kawaguchi et al. 2014), who used a non-rotational radiation source and a segmented tissue map. The μ-map was computed using a segmented MR tissue map (bone, air, other tissue), the partial path length of each tissue and the intensities of attenuated radiation that were detected from a fixed position. The partial path length was calculated from a virtual scan and the segmented MR image.

2.2 Emission-based attenuation correction

Emission-based approaches make use of the fact that the PET emission data not only contains information about the activity, but also provides information about the attenuation within the body. Therefore, these methods aim to calculate both attenuation and activity coefficients simultaneously.

2.2.1 Joint estimation of emission and attenuation

In 1979, prior to the development of integrated PET/MR systems, Censor et al. (Censor et al. 1979) presented their work on calculating attenuation and activity concentration coefficients simultaneously. Their method is based on a system of non-linear equations that describes the model of gamma-ray emission. The attenuation and activity coefficients are then calculated by iteratively refining an initial guess of both. In 1999, Nuyts et al. (Nuyts et al. 1999) carried on this model by incorporating the Poisson nature of the emission data and developed an algorithm called maximum-likelihood reconstruction of attenuation and activity (MLAA). In this method an objective function which is the sum of the likelihood and an a-priori probability about the attenuation coefficients in the human body is optimized.

However, it has been shown that without TOF the emission data are not sufficient to derive enough information about attenuation as the solution of the simultaneous estimation is not unique (Natterer & Herzog 1992). In 2012, Defrise et al. (Defrise et al. 2012) demonstrated that a unique solution for attenuation and emission, except for a constant, can be found if TOF information is available. In the same year, Rezaei et al. (Rezaei et al. 2012*a*) were able to show that TOF information can eliminate the problem of cross-talk between attenuation and activity. However, it is still necessary to include some prior knowledge, as the solution is only determined up to a scaling constant. In 2014, both Rezaei et al. (Rezaei et al. 2014) and Defrise et al. (Defrise et al. 2014) proposed to maximize the above objective function by taking not only the activity image into account, but also the attenuation sinogram, which is comprised of a set of attenuation correction factors for all lines-of-response (LORs). The so called maximum likelihood attenuation correction factors (MLACF) algorithm does not reconstruct the attenuation image, but still requires pre-knowledge about the activity or the attenuation factors.

2.2.2 Joint estimation using anatomical priors

In 2011, Salomon et al. (Salomon et al. 2011) proposed to incorporate anatomical information of the MR into the joint reconstruction of emission and attenuation for TOF PET for the lung. In this approach the local tracer concentration and the attenuation is iteratively estimated by using the segmented MR image as anatomical reference. Due to the fact that no accurate distinction between the different tissue classes is required in the beginning, the attenuation estimation is initialized by setting the attenuation coefficients in the segmented MR to the attenuation of water at 511 keV. In 2015, Mehranian and Zaidi (Mehranian & Zaidi 2015) built on this approach and constrained the MLAA algorithm to only estimate lung linear attenuation coefficients (LACs) in the segmented MR tissue map using a combination of a Gaussian and Markov random field model aiming to derive continuous lung attenuation coefficients.

2.3 Segmentation-based approaches

Segmentation-based approaches distinguish multiple distinct anatomical regions and allocate a constant pre-defined linear attenuation coefficient to each region. These linear tissue-dependent attenuation coefficients were empirically defined by the International Commission on Radiation Units and Measurements in 1989 (*Tissue substitutes in radiation dosime-*

try and measurement 1989). How many regions can be discretely classified depends on two factors: first, the MR sequence that is used for the image acquisition, and second, the method that is used to segment those regions. Segmentation-based approaches can further be divided into two sub-classes depending on whether they ignore information about bone or include it.

2.3.1 Segmentation ignoring bone

In 1984, Dixon (Dixon 1984) introduced a sequence that was able to distinguish adipose-based tissue from water-based tissue, due to the fact that the Larmor frequency of protons slightly differs for the two tissue types. Therefore, water and fat images can be derived by acquiring MR images at different echo times. In 2009, Martinez-Möller et al. (Martinez-Möller et al. 2009) segmented the human body into four tissue classes (background, lungs, fat and soft-tissue) using a 2-point Dixon sequence and by applying a threshold to both water and fat images to separate soft-tissue and fat from the background. In order to define the lung region, a connected-component analysis of the region with low MR signal in the inner part of the body was used. Some misidentified voxels were refined by applying a morphologic closing filter to the tissue-air image. In 2011, Schulz et al. (Schulz et al. 2011) proposed a method that results in a 3-class tissue segmentation (air, soft-tissue, lung) and makes use of a T1-weighted MR sequence. Both the outer body contour and the lungs were extracted using slice-wise region-growing techniques in combination with an automatically determined threshold. In 2013, Chang et al. (Chang et al. 2013) analyzed whether non-attenuated PET images (NAC-PET) can lead to a positive contribution towards PET attenuation. This method consists of three steps: segmenting the NAC-PET to initialize a first attenuation map, correcting the raw PET data for attenuation, and refining the segmentation using the corrected PET data. This iterative process first segments the outer contour of the body and the lungs before refining the lung segmentation.

2.3.2 Segmentation including bone

The previous described methods however do not allow the detection of bone, which has a significant impact on the PET quantification (Schleyer et al. 2010). Therefore, significant effort has been made to develop MR sequences aiming to extract the bone class directly and assigning a fixed attenuation value to the class.

2.3.2.1 From a T1-weighted MR sequence

In 1994, already prior to the invention of combined PET/MR systems, Le Goff-Rougetet et al. (Le Goff-Rougetet et al. 1994) attempted to extract attenuation information from brain MR images. Aiming to simplify the acquisition protocol and to reduce the dose received due to the transmission by the patient, this method segments T1-weighted MR images into bone and soft-tissue classes by using a threshold, morphologic operations and connected component analysis. In 2003, Zaidi et al. (Zaidi et al. 2003) proposed another segmentation approach for brain MR data that applies a fuzzy c-means algorithm (FCM) to T1-weighted spin-echo images in order to segment five tissue classes (air, brain tissue, skull, nasal sinuses, scalp). In 2009, Wagenknecht et al. (Wagenknecht et al. 2009) developed a three-step segmentation method making use of anatomical knowledge of the brain. In the first step, four classes (grey and white matter, cerebrospinal fluid, adipose tissue, and background) were distinguished using a neural network-based tissue classification approach. Secondly, the brain region was separated from the extracerebral region and the extracerebral region was segmented using a knowledge-based approach. Finally, the extracerebral region was segmented into multiple regions (brain tissue, extracerebral soft tissue, bone, mastoid process, and (para)nasal cavities). In 2013, Yang and Fei (Yang & Fei 2013) presented a skull segmentation method for T1-weighted MR images, in which they used a multi-scale bilateral filtering scheme to process the MR sinogram data in the Radon space.

2.3.2.2 From T1-weighted and Dixon sequences

In 2014, Anazodo et al. (Anazodo et al. 2014) combined the attenuation map acquired using a standard MR Dixon sequence for the brain with a bone mask that was generated using individual T1-weighted MR data with segmentation tools in SPM8 (`http://www.fil.ion.ucl.ac.uk`) and ICBM Tissue Probabilistic Atlases (`http://www.loni.usc.edu/ICBM/`).

2.3.2.3 From UTE sequences

The Dixon-based and other standard MR sequences do not allow the distinction between air and bone due to their low signal intensity. However, Ultrashort Echo Time (UTE) sequences are able to visualize cortical bone despite its very short T2 relaxation time (Catana et al. 2010). In 2010, UTE images were used by Keereman et al. (Keereman et al. 2010) as an input to their segmentation strategy. They acquired two UTE images at different echo times (TE) that only differed in the bone signal. Using the inverse of the T2* relaxation time

($R2* = \frac{lnUTE_1 - lnUTE_2}{TE_2 - TE_1}$) they were able to distinguish between cortical bone and soft-tissue. However, it is not as simple to differentiate soft-tissue and air due to possible artifacts and noise in the UTE images. A binary air mask of the first UTE image helped to overcome this issue. This method was further improved in 2014 by Aitken et al. (Aitken et al. 2014), who corrected the UTE image correcting for Eddy current artefacts, and in 2015 by Capello et al., who refined the R2* map segmentation.

2.3.2.4 From UTE and Dixon sequences

In 2012, Berker et al. (Berker et al. 2012) combined the advantages of both the Dixon and UTE sequences and incorporated cortical bone segmentation and water-fat decomposition. Several combinations of echoes are used to segment cortical bone, on the one hand, and separate fat and water signals, on the other hand. A predefined attenuation coefficient was assigned to the segmented bone region whereas the attenuation coefficients for each mixed water-fat voxel was calculated from the relative water fat-fraction. Hsu et al. (Hsu et al. 2013) and Su et al. (Su et al. 2015) later built on this idea by using a FCM algorithm to either segment a set of MR images or a single UTE-mDixon image.

2.3.3 Segmentation methods with subject-specific bone attenuation coefficients

Assigning predefined linear attenuation coefficients to different tissue classes, as presented in the previous section, can limit their accuracy due to the variability within the attenuating tissue. In 2015, Juttukonda et al. (Juttukonda et al. 2015) overcame this problem by thresholding R2*, fat, water and UTE images in order to segment bone, fat, soft-tissue and air. Three classes (air, fat and soft-tissue) were assigned to predefined LACs whereas segmented bone tissue was converted to attenuation values with the help of a regression model between the R2* values and bone density. Ladefoget et al. (Ladefoged et al. 2015) based their approach on the same sequences, but used a different fitting function for their model. They further added a regional mask that was defined on an atlas in their approach to separately treat complex areas with mixed air and tissue.

2.4 Atlas-based approaches

Atlas-based attenuation correction methods predict attenuation coefficients on a continuous scale by deforming an anatomical model or dataset to match the subjects anatomy using non-rigid registration. These methods usually require a number of MR/CT datasets that

form the atlas. In general, atlas-based approaches allow for the prediction of bone tissue without additional UTE imaging. Within the last ten years, several research groups focused on the development of single- and multi-atlas-based approaches.

2.4.1 Single atlas approaches

In 2007, Kops and Herzog (Kops & Herzog 2007) created an attenuation template of the brain by averaging multiple co-registered PET transmission scans before co-registering the associated MR template to the patient's MR image, resulting in a μ-map for PET AC. In 2010, Schreibmann et al. (Schreibmann et al. 2010) developed a multi-modality optical flow deformable model that co-registered a single brain CT to the target patients MR image resulting in a pseudo CT for PET AC. In 2012, Dowling et al. (Dowling et al. 2012) used a whole-pelvis MR atlas generated by manually-delineated MR scans to create subject-specific pseudo CTs. They first registered the MR atlas to the patient's MR before transforming the corresponding pseudo CT, that is in the same space as the MR atlas, using the same transformation matrix. In 2014, Izquierdo-Garcia et al. (Izquierdo-Garcia et al. 2014) used the SPM8 software to create attenuation maps based on MR and CT atlases. The atlases were created by segmenting MR images into six tissue classes and non-rigidly aligning the tissue maps before transferring the same transformation to the corresponding CT images. Pseudo CTs were generated by segmenting the subject's MR into the same six tissue classes and non-rigidly registering the tissue map to the MR atlas followed by applying the inverse transformation to the CT atlas.

2.4.2 Multi-atlas approaches

In contrast to the previously described methods, multi-atlas based approaches rely on a database of CT and MR images instead of a single atlas. Using multiple atlases aims to tackle the problem of the strong dependency of the resulting AC map on an accurate mapping between atlas and subject as well as on the representativeness of the single atlas (e.g. anatomical abnormalities). In order to build the database each CT and MR pair in the methods reported below is affinely aligned. The generation of the pseudo CT is then based on a non-rigid registration of all MR images within the database to the subject's MR. All CT images in the database are transferred to the same space by applying the resulting displacement fields to the corresponding CT images before combining the deformed CT images to the final target CT. There are several methods to fuse the deformed CT images. In 2013, Burgos et al. (Burgos et al. 2013) introduced a multi-atlas information propagation scheme

to synthesize pseudo CTs. They register all MR images in the database to the target MR image before all corresponding CTs are mapped into the same space using the same transformation. A local image similarity measure (LIS) between the mapped and target MR images is then converted into weights to generate the synthetic CT. In 2015, Sjölund et al. (Sjölund et al. 2015) generated the pseudo CT by iteratively registering them to their mean. They were able to show that the consistency of the target CT can be improved by taking the voxelwise median of the deformed CT images. In the same year Merida et al. (Mérida et al. 2015) proposed a maximum probability (MaxProb) method that analyzes the probability of each voxel belonging to a certain tissue class that has been defined by intensity thresholding. The pseudo CT was generated by calculating the average intensities on a voxelwise level of the atlases belonging to the maximum probability class. In 2017, a multi-centre study Ladefoged et al. (2017) has shown that multi-atlas propagation methods (Burgos et al. 2014) outperform methods that exploit emission data (Salomon et al. (2010), Rezaei et al. (2012*b*)) or use assigned tissue classes (Martinez-Möller et al. (2009), Catana et al. (2010)) in order to correct for photon attenuation.

2.5 Patch-based approaches

A potential problem of the multi-atlas methods is that they rely on non-rigid registration, which is time-consuming and can be unstable for subjects that do not resemble the subjects in the database, e.g., due to surgery or anatomical abnormalities. Patch-based methods try to circumvent this since atlases and target images do not have to be accurately aligned. They incorporate information of the neighboring voxels that surround a voxel of interest, which could improve finding similarities between both. In 2014, Roy et al. (Roy et al. 2014) presented a patch-based approach using intensity normalized whole head dual UTE images to generate pseudo CTs. The images are partitioned into patches and co-registered to a CT from the same subject. Patches of the reference and target MR images are matched and corresponding patches from the reference CT are combined via a Bayesian framework to create the synthetic CT image following the assumption that similar intensities in target and reference MR patches originate from the same distribution of tissues. Therefore, the corresponding reference CT patch is an approximation of the target CT patch. In 2015, Andreasen et al. (Andreasen et al. 2015) defined patches for all voxels from co-registered MR and CT brain data resulting in a database of MR and CT patches. In order to generate the pseudo CT, patches from the target MR image were extracted and an intensity-based

nearest neighbor search was run in the patch database. The synthetic CT image was then computed by identifying the K patches minimizing the squared $\mathcal{L}_2$-norm between the target and reference MR patches.

2.6 Machine learning approaches

Learning-based approaches are based on a training dataset from which the relationship between MR and CT images is learned and then applied to a target MR to synthesize a pseudo CT. These methods can generally be distinguished between supervised and unsupervised methods. In supervised methods, paired data is available, which in the context of MR to CT synthesis means MR and CT images that have been co-registered. In unsupervised methods, on the contrary, the available data does not need to be co-registered in order to generate a pseudo CT from a given MR image.

2.6.1 Supervised methods

In 2011, Johansson et al. (Johansson et al. 2011) proposed a method that uses a Gaussian mixture regression model to link the MR and CT image intensities of the training dataset. The initial method disregarded spatial information and as a result the quality of the pseudo CT was modest. As such, they later extended this method to include the spatial position of each voxel inside the head. In 2016, Huynh et al. (Huynh et al. 2016) presented a method that uses a structured random forest and auto-context model to estimate CT images. The method is also based on a training dataset of co-registered T1-weighted MR and CT images. The first step of the training partitions each MR image into a set of patches before a structured random forest is applied to directly estimate the corresponding CT patch. This initial set of predictions was refined using an auto-context model. The final pseudo CT is then generated by combining all predicted CT patches.

Since 2016, there has been a shift of emphasis in the field of PET/MR attenuation correction towards deep learning approaches that have proved to be a powerful tool in the MR to CT image translation outperforming state-of-the-art multi-atlas-based methods (Burgos et al. 2013). Image synthesis can be considered as a regression problem, where for every input pixel a corresponding output pixel is required. Many deep learning methods employ convolutional neural networks (CNN) that are able to capture the contextual information between two image domains (as between MR and CT) in order to translate one possible representation of an image into another. One of the first approaches was introduced by Nie

et al. (Nie et al. 2016) and makes use of a 3D deep learning-based method to predict a pseudo CT. The relationship between MR and CT images is learned by a 3D fully convolutional neural network (FCN) that preserves the neighborhood information better than a CNN. Differing to a CNN, the proposed FCN predicts the pseudo CT in a patch-by-patch manner. Each patch is fed into the network that consists of three 3D convolutional layers constructing 32, 64 and 32 feature maps respectively. The output layer then generates pseudo CT image patches by applying just one filter of size $3 \times 3 \times 3$, which are combined to create the entire pseudo CT image. They tested their network on a pelvic dataset and were able to show that it outperforms other three state-of-the-art methods (atlas-based method, structured random forest-based method, structured random forest and auto-context model). In 2017, Han presented a 2D deep CNN that directly learns a mapping function to translate a 2D MR image slice into its corresponding 2D CT image slice (Han 2017) closely following the U-Net architecture, which has gained recognition in the deep learning community due to its strong performance in the field of image segmentation Ronneberger et al. (2015). The original U-Net architecture can be seen in Fig. 2.1. Nie et al. further proposed a pseudo CT synthesis method that utilizes a fully-connected neural network with an adversarial learning strategy (Nie et al. 2018). A second network discriminates the output of the fully-connected network and can urge it to look as similar to the real CT as possible. In 2018, Emami et al. introduced a generative adversarial synthesis framework that uses a deep residual network as image generator (Emami et al. 2018).

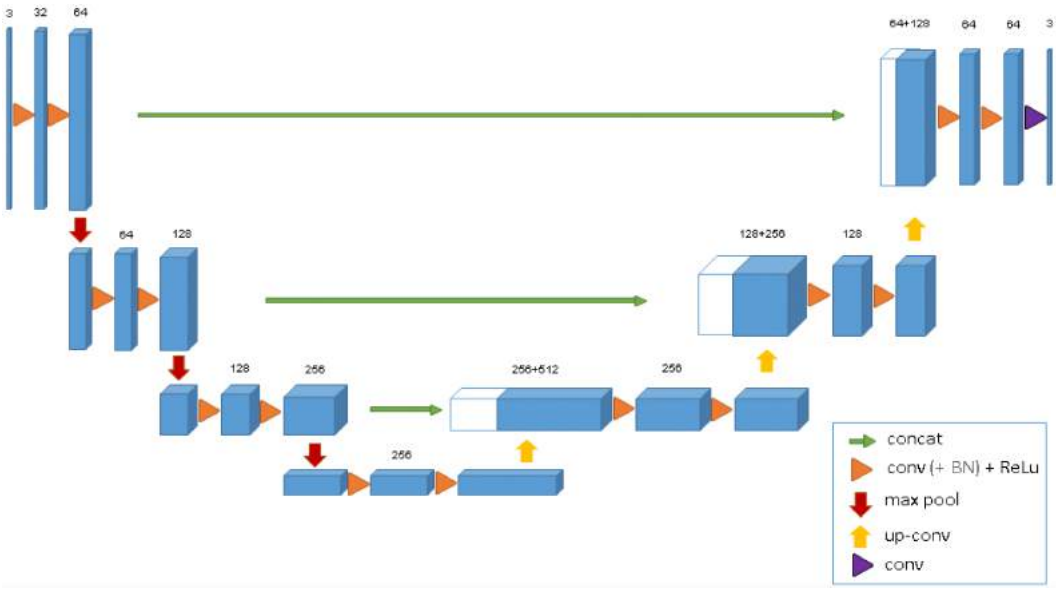

Figure 2.1: Original 3D U-Net architecture from Çiçek et al. (Çiçek et al. 2016).

2.6.2 Unsupervised methods

With the emergence of the cycleGAN in 2017 (Zhu et al. 2017), a large amount of work has been conducted in the field of unsupervised pseudo CT synthesis. Unsupervised learning scenarios disregard the need of paired data and the ill-posed $\mathcal{L}_2$-loss, commonly used in supervised MR to CT synthesis. A generative adversarial network (GAN) is a machine learning framework that consists of a generator and a discriminator. The generator generates new data with the same statistics as the training data from an initial estimate, i.e. noise. The discriminator aims to distinguish between real (from the domain) or fake (generated) images. The cycleGAN builds on the GAN framework idea. It consists of two pairs of simultaneously trained generators and discriminators. The first generator uses images from domain A to translate them into domain B, and the second generator works vice versa (i.e. images are translated from domain B into domain A). Additionally, the framework makes use of a concept called *cycle consistency*, which means that the first generator translates the input image A into an ouput image B that when used as input to the second generator generates the original image A. Wolterink et al. (Wolterink et al. 2017) presented a CNN in their work that minimizes an adversarial loss to learn a mapping fuction between MR and CT. This adversarial loss encourages the pseudo CT to be indistinguishable from the ground truth CT. An additional CNN aims to assure that the pseudo CT corresponds to the actual input MR image. However, using a cycleGAN alone for pseudo CT synthesis does not automatically ensure that pseudo CT and ground truth CT are structurally consistent. This means that the reconstructed MR image is almost identical to the input MR, however, the pseudo CT is significantly different from the ground truth CT. Therefore Yang et al. (Yang et al. 2018) proposed a cycleGAN containing structural constraints by minimizing an additional structural consistency loss. These methods have demonstrated that it is possible to generate pseudo CT images from unpaired MR data of the brain. However, there have been multiple other approaches to synthesize pseudo CT images of other body parts. Zhang et al. presented a cycle-consistency adversarial network for cardiovascular pseudo CT volumes (Zhang et al. 2018), Huo et al. proposed a network for MR to CT synthesis and segmentation of the spleen (Huo et al. 2018) and Hiasa et al. added a gradient consistency loss to the original cycleGAN in order to synthesize pseudo CT images from musculoskeletal MR images (Hiasa et al. 2018).

2.7 Overview of advantages and disadvantages of various AC methods

Although many different approaches have been proposed, the field of PET/MR attenuation correction remains an important topic of discussion in medical imaging research. To summarize what work has been done so far, a list of the advantages and disadvantages of the various methods for PET/MR attenuation correction is shown in Table 2.1.

2.8 Discussion

Transmission-based methods were the first to be developed for the purpose of PET attenuation correction and are a promising approach due to their ability to directly result in a map of linear attenuation coefficients. For a long time, attenuation maps based on transmission scans have been considered as the gold standard, however, they were developed long before simultaneous PET/MR systems, which is why in practice it is difficult, if not impossible, to implement a rotating transmission source inside the small bore of a PET/MR scanner. Furthermore, the radiation dose received by the patient would increase, which is counterintuitive when thinking of the benefit of the reduced radiation of a PET/MR scanner.

Emission-based attenuation correction methods have also shown promising results over the last years, benefiting from the fact that no task-specific MR sequence needs to be acquired. Studies have shown promising results, but require more validation on patients in order to fully assess their capabilities to accurately correct for attenuation in clinical PET data. Moreover, these methods are limited as most approaches require TOF information, which is not available in all PET/MR systems currently on the market, thus making those methods clinically less attractive.

Segmentation-based approaches distinguish multiple distinct anatomical regions and allocate a pre-defined linear attenuation coefficient to each region. The success of these methods therefore relies on the quality of the MR acquisition. Segmentation-based methods were the first to be implemented in commercial PET/MR scanners, however, until this point, they struggle to provide sufficient information about bone, which is essential when correcting for PET attenuation. Bone is the tissue class that attenuates photons the most, wherefore small changes in misclassified bone tissue can result in large quantification errors in the reconstructed PET image. In addition, the attenuation coefficients assigned to each tissue are fixed and therefore further limit the accuracy when used for attenuation correction.

Method	Advantages	Disadvantages
Transmission-based (PET)	Results directly in LACs; determination of LACs of any object in FOV; no coil template needed	Increased dose and additional source; implementation of rotating source difficult
Emission-based (PET)	No need for task-specific MR sequence	Limited to tracers with distributed uptake; currently requires TOF information
Segmentation-based	Fast; individual patient data; reference data not needed	Robustness depends on anatomical assumptions; additional acquisition time for bone signal; fixed AC value per tissue, so inadequate for the lung; truncated FOV
Atlas-based	Continuous pseudo CT values based on population	Morphological abnormalities; multiple non-rigid registrations; truncated FOV; difficult for whole-body
Patch-based	Do not require an accurate alignment between atlases and target; no non-rigid registration needed	Carefully designed patch search to guarantee reasonable run time
Learning-based	Fast as no non-rigid registration needed; ability to capture complex non-linear mapping from input to output space	Need of images with sufficient contrast to distinguish between air, bone, and soft-tissues; highly dependent on size of database; computationally expensive to train the model

Table 2.1: Advantages and disadvantages of approaches for attenuation correction in PET/MR.

Atlas-based methods are able to generate continuous pseudo CT values and have shown state-of-the-art results for many years. However, given the multiple non-rigid registrations often needed in such methods, they tend to be slow. Moreover, they are limited to anatomical features present in the atlas database and thus not able to model any anatomical abnormalities. This can be particularly concerning in clinical practice where PET/MR is often used as part of oncological examinations where abnormalities are expected. Inaccurate PET reconstructions could potentially lead to missing crucial information such as metastatic activity. Due to their state-of-the-art performance a multi-atlas propagation method was chosen for comparison in this thesis for experiments performed on the skull. Since it is not inconsequential to generate a perfectly co-registered database for such methods, experiments on the full body (see chapter 5) were only compared to other deep learning-based methods.

While a lot of progress has been made in the field of PET/MR attenuation correction, no universal method has been established that is robust enough to be routinely used in clinical practice. Often, patients present unique anatomical abnormalities that state-of-the-art methods like multi-atlas propagation methods are not able to grasp. In the case of neurological studies, for example, the resulting imperfect attenuation correction can lead to a strong bias in the reconstructed PET distribution, which can result in a wrong diagnosis. Therefore, it is essential to develop new methods to ensure accurate attenuation correction of PET/MR data in the brain. While the progress in PET/MR attenuation correction has been significant in recent years, there has been little progress in whole-body PET/MR applications. However, whole-body PET/MR imaging has a promising future ahead, e.g. in the field of oncology, where accurately reconstructed whole-body PET images are needed to detect and monitor metastatic activity.

Some groups have proposed to directly synthesize attenuation corrected PET images from MR images directly circumventing the interim step of synthesizing a pseudo CT image. Sikka et al. adapt the original 3D U-Net architecture to a global and non-linear cross-modal approach that estimates PET images from MR images directly (Sikka et al. 2018). Hwang et al. combine the traditional maximum-likelihood reconstruction of activity and attenuation (MLAA) method (Rezaei et al. 2012*b*) with deep learning in order to overcome the limitations of MLAA (Hwang et al. 2018). Yaakub et al. propose a method to synthesize pseudo-normal PET images from MR images in a generative manner in order to identify regions of hypometabolism in PET images of epilepsy patients (Yaakub et al.

2019). However, a general difficulty that direct MR to PET synthesis methods face is the fact that the two imaging modalities depict inherently different information: MR describes anatomical information whereas PET is a functional imaging technique. Additionally, PET reconstruction depends on the dose of the injected tracer that is taken into account in the reconstruction process as a parameter. While both MR and CT images are also very different in the information that they provide, they do both descibe anatomical features, thus making the translation task easier.

The MR to CT translation task is not only crucial for attenuation correction, but also plays an important role in radiotherapy treatment planning. Radiotherapy aims to deliver an optimal dose of radiation to the cancerous area while minimizing the dose received by healthy tissues. Knowledge about different tissue attnuation properties is necessary to determine the optimal dose distribution for attacking the cancerous area. Therefore, in clinical practice, a CT scan is acquired. However, the soft tissue contrast in CT images is not strong, which can cause large variations in segmenting the tumor, both in the brain and in the entire body. It is desirable to exploit the excellent soft-tissue contrast from MR images in order to be able to delineate tumors and organs at risk more accurately. Similarly to the problems present in MRAC, tissue attenuation coefficients are not easily estimated from MR images. Therefore, many groups have attempted to synthetically create CT images from MR images specifically for radiotherapy. In 2014, Korhonen et al. (Korhonen et al. 2014) presented a method where they manually segmented bones from the MR image before converting the image into HUs. Jonsson et al. (Jonsson et al. 2015) proposed a method where they utilize a Gaussian mixture regression model that links MR and CT intensities. Common approaches to generate a CT from an MR image in the field of MR only radiotherapy planning are atlas-based. Earlier methods rely on a single atlas (Dowling et al. 2012), while newer methods make use of a database of multiple atlases (Sjölund et al. 2015, Gudur et al. 2014, Uh et al. 2014). They mainly differ in the fusing technique applied to the registered CT images (voxelwise median, probabilistic Bayesian framework, arithmetic mean process, pattern recognition with Gaussian process). More recently, deep learning has been employed for MR-only radiotherapy treatment planning. While some groups used fully connected neural networks (Zhao et al. 2018), others attempted to solve the problem with GANs (Maspero et al. 2018). The field of MRAC and MR-only radiotherapy are similar in many aspects, potentially benefitting from each other, while aiming to improve different

applications. In radiotherapy MR and CT images of the same patient are acquired routinely for dose estimation, which could be a potential data source for deep learning-based MRAC methods that rely on a large amount of data for training. Vice versa, deep learning-based CT synthesis methods have the potential to improve radiotherapy treatment planning such that an additional CT acquisition becomes redundant.

Since this project's inception in 2016, only a few attempts have been made to solve the PET attenuation correction task with deep learning methods. The majority of them focused on the skull, some attempted to synthesize pseudo CT images of the pelvis, however, no group had attempted to synthesize whole-body pseudo CT images. At this point, deep learning-based methods had gained a lot of popularity in the field of computer science, showing promising performance in the fields of image classification, segmentation and synthesis. This led to the project aim of developing deep learning methods for medical image synthesis.

Chapter 3

Deep learning in medical imaging

The first contribution to this work on PET/MR attenuation correction is a CNN that makes use of the idea of an algorithm called *Boosting* known from classic machine learning (Schapire 1990). The idea of the boosting algorithm is to combine a sequence of weak learners that in their entirety build a strong learner. This way, each model aims to compensate the weaknesses of its predecessors. In the context of neural networks this means concatenating multiple CNNs, each representing a weak learner, that when trained all together build a strong learner that predicts a more accurate pseudo CT. The method can also be seen as a form of residual learning, where the residuals of an initial prediction are minimized by additional learners further down the stream. The development of the final method followed multiple steps starting with predicting pseudo CT images as a classification problem through the regression of continuous voxel values via direct and recursive image synthesis to the final algorithm called *Deep Boosted Regression (DBR)* published in (Kläser et al. 2018).

All contributions to this thesis have been implemented and carried out using *NiftyNet*, which is a TensorFlow-based deep learning framework tailored for medical imaging (Gibson et al. 2018).

3.1 Experimental dataset

The experimental dataset used in this chapter consisted of 20 pairs of brain MR, CT and ^{18}F-FDG PET images. All 20 subjects were scanned on a 3T Siemens Magneton Trio scanner and T1-weighted (TE/TR/TI,2.9 ms/2200 ms/900 ms; flip angle 10°; voxel size $1.1 \times 1.1 \times 1.1$ mm^3) and T2-weighted (TE/TR, 401 ms/3200 ms; flip angle 120°; voxel size $1.1 \times 1.1 \times 1.1$ mm^3) volumetric scans were acquired. PET/CT imaging was performed on a GE Discovery ST PET/CT scanner providing CT images (voxel size $0.586 \times 0.586 \times 2.5$ mm^3,

120 kVp, 300mA) and reconstructed PET images (voxel size $1.95 \times 1.95 \times 3.27$ mm^3). For each subject MRs and CTs were affinely aligned using a symmetric approach (Modat et al. 2014) based on Ourselin et al. (Ourselin et al. 2001) followed by a fully affine registration in order to compensate for possible gradient drift in the MR images. A very low degree of freedom non-rigid deformation (i.e. low resolution control point grid with spacing of 7.5 mm along each axis) was performed afterwards in order to compensate for different neck positioning before implementing a second non-linear registration, using a cubic B-spline with normalized mutual information (Modat et al. 2010). Each volume had $301 \times 301 \times 153$ voxels with a voxel size of approximately 1 mm^3. Both CT and MR images were rescaled to be between 0 and 1 for increased training stability. For evaluation purposes two masks were extracted, a head mask from the CT and a brain mask from the T1-weighted MR image. The head mask was generated by thresholding the CT at -500 HU thus excluding the background from the performance metric analysis. In order to evaluate the performance of the pseudo CT when used for PET attenuation correction, an additional brain mask was extracted from the T1-weighted MR image to exploit the radionuclide uptake in the brain region only. Registration quality was carefully assessed manually by multiple specialists for each subject. The success of the supervised learning methods highly depends on the registration quality of the MR/CT database. Even small inaccuracies in the registration can influence the training and subsequently lead to underestimation of the attenuation map, wherefore careful registration quality assessment is essential. The data were split into 70% training, 10% validation and 20% testing data for all methods.

3.2 CT image synthesis as a segmentation problem

At the beginning of this project in 2016, few attempts had been made to synthesize CT from MR images using deep learning. The development of the first published work (Kläser et al. 2018) followed several stages. The first approach used in order to synthesize a pseudo CT from an MR image was to consider the image-to-image translation task as a multi-class segmentation problem instead of a regression task that predicts CT images in a continuous scale. Therefore, all CT images have been transformed into label images containing 26 classes. In order to generate these label CTs, the original continuous CTs were linearly rescaled to be in a range of 0 and 26. Afterwards all values within the rescaled continuous CTs were discretized. An example can be seen in Fig. 3.1. The label CT consisting of 26 classes is pictured on the left whereas the original CT image with continuous values in the

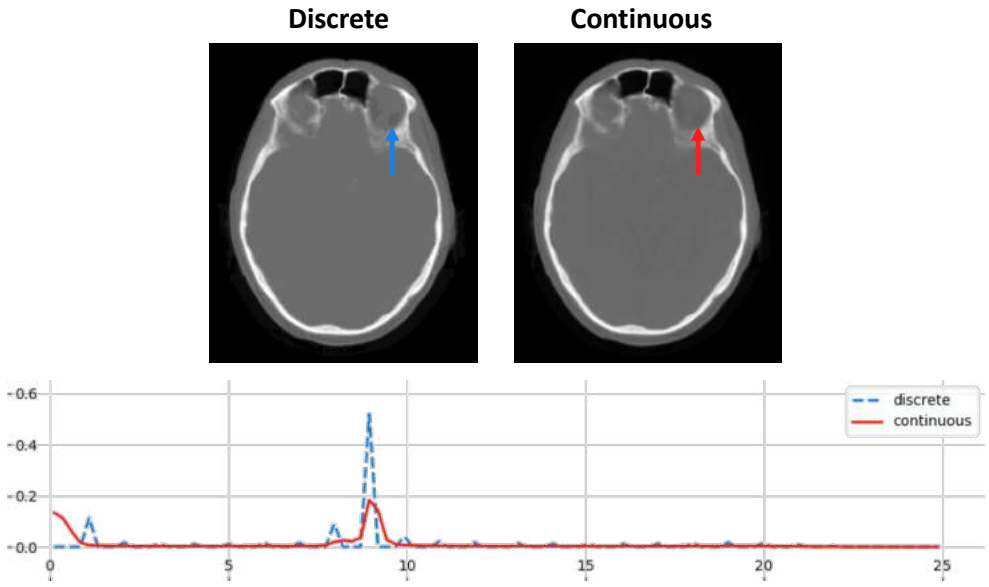

Figure 3.1: Discrete CT with 26 classes (left) and CT with continuous pixel values (right) and corresponding histograms (bottom). The main difference is pointed out by arrows.

range of 0 to 26 HU can be seen on the right hand side of the figure. It is evident that 26 classes are sufficient to approximate a realistic looking CT. The corresponding histograms are shown at the bottom of Fig. 3.1. The majority of pixel values within the continuous CT are in a range between 0 and 1 and 7 and 10 with few outliers outside these ranges. Looking at the histogram of the discrete CT labels, the biggest difference can be observed in the background as the probability of a pixel to be closer to 1 is higher than to be 0. The majority of bone is classified as a discrete label of value 7, which correlates with the distribution of the real CT values, however, the peak for the discrete CT is sharper explaining a less smooth image. The arrows show an obvious difference between the discrete and the continuous CT images.

3.2.1 Implementation details

In order to train the proposed multi-class segmentation problem, a high resolution compact network architecture known as HighRes3DNet presented by Li et al. (Li et al. 2017) was adopted. This network was introduced for the purpose of volumetric image segmentation and is very efficient in learning 3D representation from large-scale image data. It consists of 20 convolutional layers with kernel size $3 \times 3 \times 3$ that encode low-level image features.

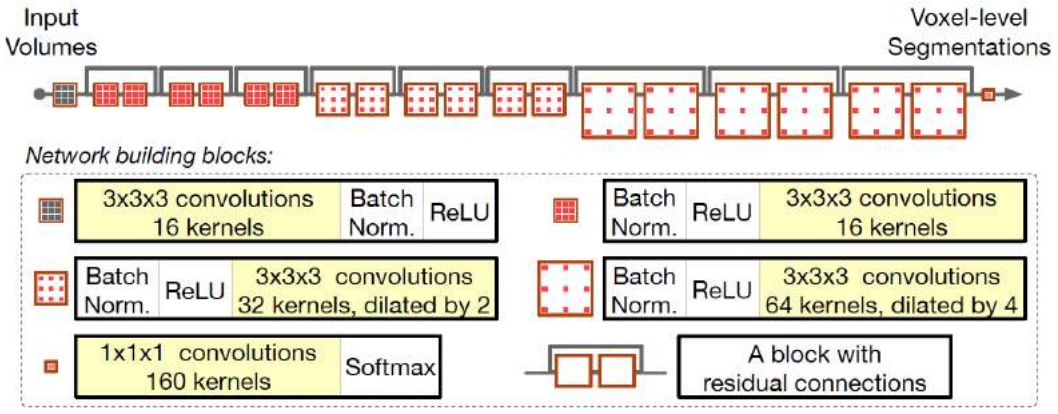

Figure 3.2: Original HighRes3DNet architecture from Li et al. (2017). Two crucial building blocks of this network are a) dilated convolutions with gradually increasing dilation factors in order to capture features at multiple scales and b) residual connections enabling identity mapping such that features from different scales can be connected. The network ensures that the spatial resolution of the input image is kept the same throughout the network.

Mid- and high-level image features are captured within the following convolutional layers with kernels that are dilated by a factor of two and four respectively, preserving the spatial resolution of the input image throughout the network. Convolutional layers are grouped into pairs of two, and residual connections are added that enable an identity mapping so that both parameters and computational cost are minimal as shown by He et al. (He et al. 2015*a*). HighRes3DNet learns 3D representations of the data that are then mapped to the domain of CT images through a series of 1D convolutions with non-linear activation functions. The model can be trained end-to-end to directly generate the 3D pseudo CT images. The original network architecture, referred to as HighRes3DNet, can be seen in Fig. 3.2.

The network was trained with the original proposed settings, also referred to as hyper-parameters. In the scope of this thesis, the HighRes3DNet architecture has been adapted for the purpose of synthesizing pseudo CT images from MR images, however, the base structure was kept the same.

3.3 CT image synthesis using HighRes3DNet

Attempting to synthesize pseudo CT images as part of a multi-class segmentation task provides a realistically looking approximation of a CT image, however, it is desirable to generate pseudo CT images that result in an attenuation map with continuous attenuation factors. Therefore, the initial method must be adapted in order to solve a pixel-wise regression problem instead of a multi-class segmentation problem. HighRes3DNet ensures that the spatial resolution of the input image is kept constant throughout the network, which proves useful

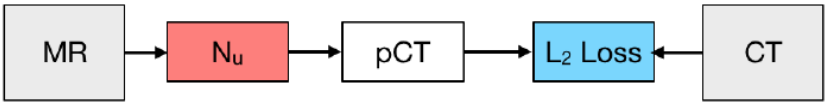

Figure 3.3: Initial network architecture for solving the CT Image synthesis task as a regression problem. The MR is fed into a network N_u and a pseudo CT (pCT) is generated by minimizing an $\mathcal{L}_2$-loss (here: RMSE) between real CT and pCT. N_u can be any network architecture suitable for image-to-image translation. Here, HighRes3DNet is used.

when regressing corresponding MR and CT patches. In order to use HighRes3DNet for direct image-to-image translation, multiple hyperparameters were optimized as well as the image value range of input and output images (see section 3.1). Figure 3.3 shows a simplified network diagram for synthesizing pseudo CT images with continuous pixel values from MR images. In the scope of this thesis, direct pseudo CT synthesis with HighRes3DNet is used as the baseline for all experiments.

3.3.1 Implementation details

The HighRes3DNet was trained on images that were rescaled to values between 0 and 1 and a Parametric ReLU activation (PReLu) was used that proved to be more robust during training. PReLUs have a small positive slope for negative values, instead of altogether zero as in the traditional ReLU. The slope is a trainable variable itself such that it can be learned along other network parameters. PReLUs have shown to improve network training at nearly zero extra computational cost (He et al. 2015*b*). In the training stage, the data were randomly sampled into subvolumes of size $56 \times 56 \times 56$ pixels due to a limited GPU memory budget. The subvolumes were then augmented by randomly rotating each of the three orthogonal planes on the fly by an angle in the interval of [-10°,10°]. The MR data was likewise randomly scaled by a factor between 0.9 and 1.1. Rotation and scaling are known methods to augment data, a commonly used way to artifically increase the size of the training dataset when only a small amount of data is available. Patches were sampled uniformly from the input images. The network was trained from scratch on a single NVIDIA Titan X GPU using the Adam optimization method. The model was trained with a learning rate of 0.001 until convergence, where convergence is defined as a sub 5% change in the loss value over a period of 5000 iterations. In this work, the root mean square error (RMSE) was chosen for loss minimization, a process that tries to bring the residuals to the smallest possible value. The RMSE is often referred to as a form of $\mathcal{L}_2$-loss, where the $\mathcal{L}_2$-norm is a measure of distance between two vectors, equivalent to the residuals. The RMSE is defined

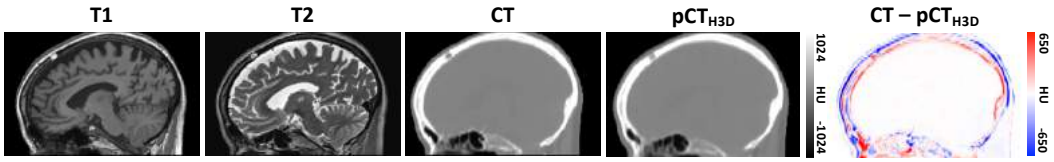

Figure 3.4: Example pseudo CT generated with HighRes3DNet only and corresponding error alongside input T1- and T2-weighted MR images and ground truth CT.

as the square root of the mean square of the error $(y - \hat{y})$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{3.1}$$

3.3.2 Results

Figure 3.4 shows an example pseudo CT and the corresponding residuals when generated with HighRes3DNet on the hold-out test set. It can be seen that the pseudo CT visually looks similar to the ground truth CT. The main source of error comes from an intensity underestimation within the skull and an overestimation in the nasal cavities, which is equivalent to assigning any intensity other than air (-1024HU).

3.4 Recursive CT image synthesis

In order to improve the performance of the previous presented network, which only uses MR input images to output pseudo CTs, a recursive network architecture was introduced and can be seen in Fig. 3.5.

The objective is to simplify the training by also inputting the previous generated pseudo CTs. At first, the network generates an initial pseudo CT containing a relatively high error

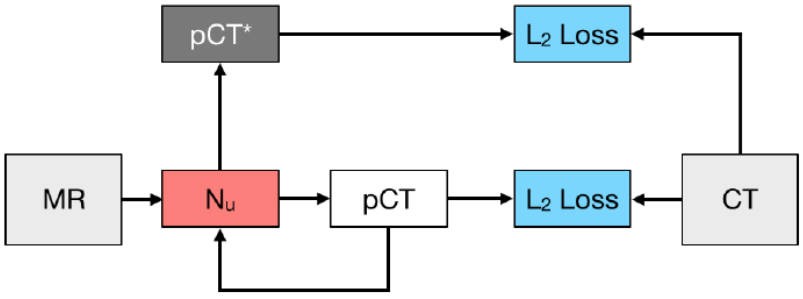

Figure 3.5: Recursive network architecture for pseudo CT synthesis. MR images are fed into a network N_u and an initial pseudo CT (pCT) is synthesized by minimizing an $\mathcal{L}_2$-loss (here: RMSE) between pCT and real CT. The pCT is then fed back into the same network in order to synthesize an improved version pCT*, also by minimizing another $\mathcal{L}_2$-loss between pCT* and real CT.

rate just like the framework presented in the previous section 3.3. In order to reduce this error, the initial pseudo CT is fed back into the same network. The network then optimizes its weights based not only on the MR images, but also on the previous generated pseudo CTs. Since MR and CT images vary quite differently in the information they contain, it is much easier for the network to learn the context between images of the same modality, the initial pseudo CT can be seen as a prior for the network. Therefore the network minimizes two $\mathcal{L}_2$-losses: 1) the error between initial pCT and true CT and 2) the loss between true CT and an updated version of the pCT (denoted as pCT*) that was corrected for its residuals.

3.5 Deep Boosted Regression

Up to this point, all work has been focused on adapting existing methods for the application of MR to CT synthesis. These stepping stones led to the first main contribution to this thesis, namely Deep Boosted Regression (DBR), presented in (Kläser et al. 2018). The method also utilizes the HighRes3DNet as it has shown to generate realistic pseudo CT images while having an efficient parameter count and large receptive field.

The aim of the proposed image synthesis approach is to find a mapping from the domain of T1- and T2-weighted MR input images to the domain of CT images. This mapping can be formulated as

$$\mathbb{R}^{T_1,T_2} \rightarrow \mathbb{R}^{CT},$$

which is a mapping from $y \hookleftarrow f(x)$, where f is a function that maps input $x \in \mathbb{R}^{T_1,T_2}$ to $y \in \mathbb{R}^{CT}$. This mapping function is highly nonlinear, and can be approximated by a composition of simpler functions with parameters ϕ, of the form $\tilde{y} = f^{(n)}(f^{(n-1)}(...(f^{(2)}(f^{(1)}(x,\phi_1),\phi_2),...),\phi_{n-1}),\phi_n)$. In a supervised learning context, these parameters ϕ are determined by minimizing a loss function that aims to minimize the residuals between the predicted CT, $\tilde{y}$, and the true CT, y,

$$\mathcal{L}_2 = ||y - \tilde{y}||_2.$$

However, the large number of functions and parameters ϕ creates computational and optimization challenges. To avoid this, the problem is reformulated as a boosting model, whereby the output of each function $f^{(n)}$ aims to approximate y. If $\tilde{y}_1 = f^{(1)}(x,\phi_1)$, then subsequent functions can be seen as a form of corrective learning, as $\tilde{y}_2 = f^{(2)}(\tilde{y}_1,x,\phi_2)$.

Thus, the model above can be rewritten as

$$\tilde{y} = f^{(n)}(f^{(n-1)}(...(f^{(2)}(f^{(1)}(x,\phi_1),x,\phi_2),...),x,\phi_{n-1}),x,\phi_n).$$

It is important to note that this corrective learning model introduces more parameters for every corrective function f, which can result in model overfitting and increases the difficulty of the optimization process. To circumvent this problem, a single corrective function $f^{(c)}$ is introduced, equivalent to sharing parameters between functions $f^{(2)}$ to $f^{(n)}$. This corrective function is applied recursively after an initial approximation of $\tilde{y}$ given by $f^{(1)}$. The recursion can be defined as

$$\tilde{y}_k = \begin{cases} f^{(1)}(x \mid N_1) & if\ k = 1 \\ f^{(c)}(x,\ \tilde{y}_{k-1} \mid N_c) & if\ k > 1 \end{cases}$$

where a function with parameters N_1 synthesizes $\tilde{y}_1$ from an input MR image x, at iteration $k = 1$. For $k > 1$, a corrective function, with parameters N_c, maps the previous prediction $\tilde{y}_{k-1}$ and the input MR images x to a better approximation of the true CT y. Finally, to ensure that the function's parameters can be optimized, the loss function is adapted to

$$Loss = \sum_{k=1}^{n} \|\tilde{y}_k - y\|^2 .$$

thus providing a form of deep supervision by introducing gradients for each function f.

The functions described above are approximated by two separate CNNs, both following the network architecture of HighRes3DNet. The proposed network architecture is illustrated in Fig. 3.6. The first network N_1 is trained to synthesize an initial pseudo CT taking both T1- and T2-weighted MR images as inputs. This first pCT is passed to a second network N_c that learns the residuals between pCT and the real CT. Therefore the weights of N_c depend on the output of N_1, but not vice versa. An improved pCT is then generated by adding the residuals to the initially synthesized pCT, which is then again fed back into N_c in order to update the weights of the network. By sharing the parameters of N_c no additional parameters are introduced to the network keeping computational complexity within limits and thus enabling a more generalized model even if only a limited number of training datasets are available. This recursive cycle can be repeated for k iterations, however, the number of iterations is limited to avoid overfitting. The proposed DBR approach exploits

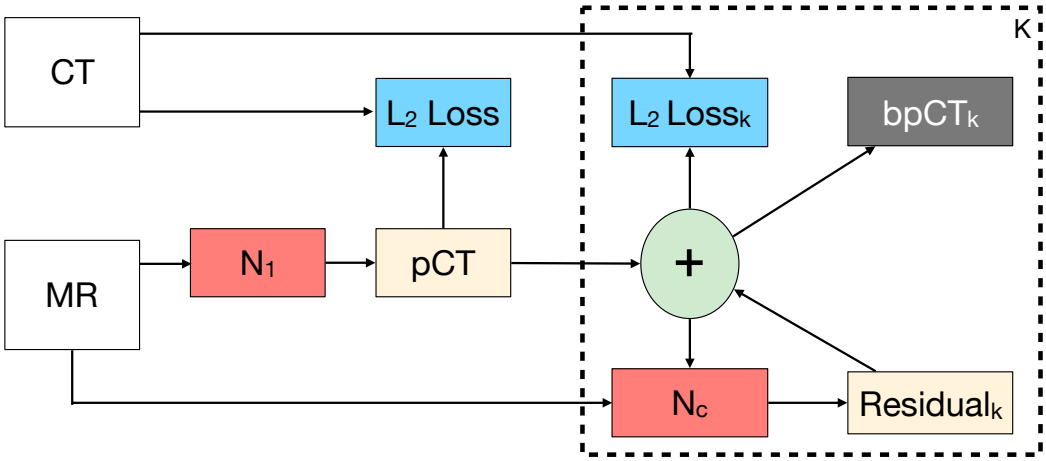

Figure 3.6: Framework of proposed Deep Boosted Regression method. MRs are fed into a first network N_1, an initial pseudo CT (pCT) is synthesized by minimizing the loss between pCT and original CT. Within the space K, residual learning is performed, where the residuals are added to pCT and fed into a second network N_c, wherefore the "+" illustrates an accumulator. A second loss is introduced minimizing the difference between ground truth CT and updated pCT. The final output is an error boosted pCT (bpCT). The number of residual learning cycles (K) is limited to avoid overfitting and was determined empirically (here, K=4).

the advantages of the recursive boosting model and is therefore independent of the choice of the cost function.

3.5.1 Implementation details

The proposed DBR network was trained on images that were rescaled to values between 0 and 1 and a PReLu was used that proved to be more robust during training. During training, the data were randomly sampled into subvolumes of size $56 \times 56 \times 56$ pixels that were augmented by randomly rotating each of the three orthogonal planes on the fly by an angle in the interval of [-10°,10°]. The MR data was also randomly scaled by a factor between 0.9 and 1.1 and patches were sampled uniformely from the input images. The network was trained from scratch on a single NVIDIA Titan X GPU using the Adam optimization method. The model was trained with a learning rate of 0.001 until convergence, where convergence is defined as a sub 5% change in the loss value over a period of 5000 iterations. Both networks N_u and N_c minimized the RMSE-loss.

3.5.2 Results

Figure 3.7 demonstrates an example pseudo CT and the corresponding residuals when generated with the proposed DBR method on the hold-out test set. It can be seen that the pseudo CT looks similar to the ground truth CT. Similarly to the results of the pseudo CT

images generated with HighRes3DNet (Fig. 3.4), the main source of error comes from the bones within the skull. Bone intensities are partly underestimated whereas other parts are overestimated. In total, the error is lower than when synthesizing pseudo CT images with HighRes3DNet only.

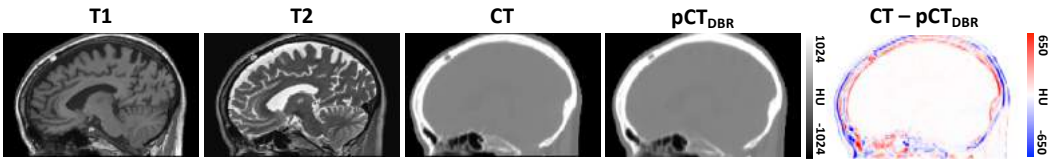

Figure 3.7: Example pseudo CT generated with proposed DBR method and corresponding error alongside input T1- and T2-weighted MR images and ground truth CT.

3.6 Comparison to state-of-the-art CT synthesis

In order to evaluate the performance of the proposed DBR method, the synthesis results were compared to three baseline methods: a multi-atlas information propagation method, a popular deep learning network (U-Net) and the HighRes3DNet (see section 3.3) that was used within the boosting framework.

3.6.1 Multi-atlas propagation

The first method that was used for comparison is a state-of-the-art multi-atlas propagation method by Burgos et al. (Burgos et al. 2013). The framework is shown in Fig. 3.8. This method relies on a well-registered database of paired MR and CT images. In order to generate a pseudo CT from any given MR image, they first register all MR images within the database to the target MR image. In the subsequent step all corresponding CT images are mapped into the same space using the same transformation as from database MR image to target MR image. A local image similarity measure (LIS) between the mapped and target MR images is then converted into weights to generate the synthetic CT.

An example of the CT synthesis results on the hold-out test set using multi-atlas information propagation and the corresponding error can be seen in Fig. 3.9. The method is able to generate a realistic looking CT with low error within the cranial vault. In general, the highest error can be observed within the bones of the skull and in the nasal cavities. It can be seen that the method fails to reconstruct the epidermoid cyst in the subject's skull.

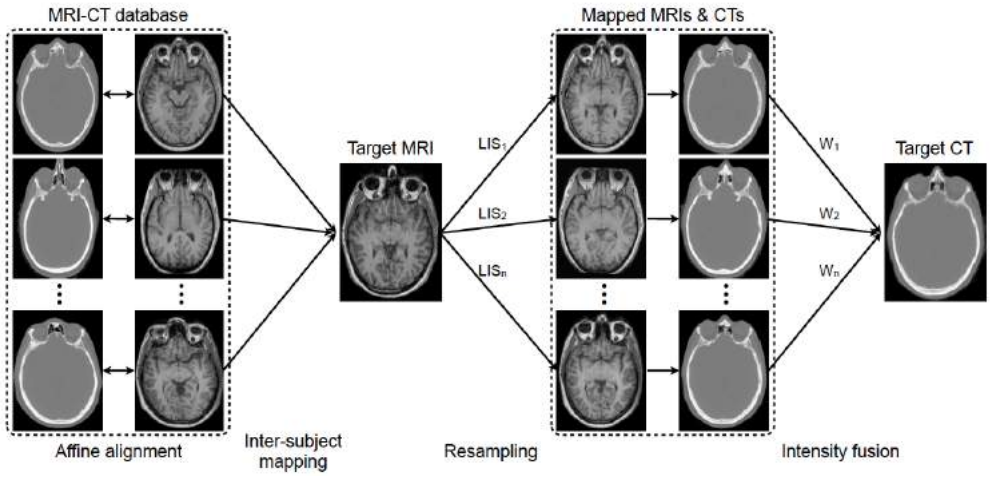

Figure 3.8: CT synthesis from a CT-MR database. All MRs within the database are mapped to the target MR before corresponding CTs are mapped to the target using the same transformation. A local image similarity measure (LIS) between the mapped and target MR images is then converted into weights to generate the synthetic CT (Burgos et al. 2014).

3.6.2 U-Net

The second method that was chosen as a baseline is 3D U-Net (Çiçek et al. 2016). The U-Net architecture gained popularity after reaching superior performance in multiple segmentation tasks. Its architecture can be seen in Fig. 2.1. The U-Net consists of two distinct paths: an analysis path (downstream) and a synthesis path (upstream). Each of the two paths consists of four blocks. In the analysis path, each block is built of two $3 \times 3 \times 3$ convolutions, one ReLu and one $2 \times 2 \times 2$ downsampling layer, whereas blocks in the synthesis path consist of a $2 \times 2 \times 2$ upsampling layer, two $3 \times 3 \times 3$ convolutions and one ReLu. Additionally, skip-connections from layers of equal resolutions are implemented. The last layer is a $1 \times 1 \times 1$ convolution in order to map the features into the number of output labels (here, into continuous values of the CT image domain).

An example for synthesis results on the hold-out test set is shown in Fig. 3.10. It can be seen that the network is able to generate a realistic looking pseudo CT image. The main

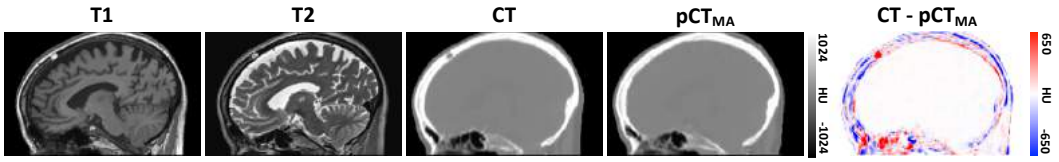

Figure 3.9: Example pseudo CT generated with state-of-the-art multi-atlas propagation method and corresponding error alongside input T1- and T2-weighted MR images and ground truth CT.

source of error arises from within the skull region. The network tends to underestimate the intensities within the skull. Further errors can be observed in the nasal cavities. It is important to note that the model is able to reconstruct the epidermoid cyst in the subject's skull.

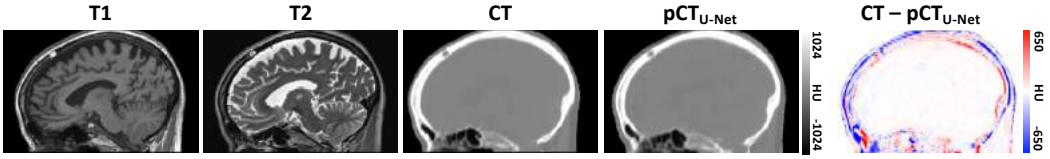

Figure 3.10: Example pseudo CT generated with U-Net and corresponding error alongside input T1- and T2-weighted MR images and ground truth CT.

3.7 PET reconstruction

The objective of this project is to synthesize pseudo CT images from MR images for PET/MR attenuation correction. Therefore, it is important to evaluate the effect of the synthesized pseudo CT images when used as attenuation maps for PET reconstruction. All PET images were reconstructed using NiftyPET, an open-source package for high-throughput PET image reconstruction (Markiewicz et al. 2018). Access to the raw PET data was not granted meaning the following simulation was performed to reconstruct PET images (see Fig. 3.11): firstly, attenuation factor sinograms were generated by forward projecting the μ-map transformed versions of each pseudo CT. Secondly, simulated emission sinograms were acquired using a similar forward projection applied to the original PET images. The simulated emission sinograms are then attenuated through element-wise multiplication with the attenuation factor sinograms. In the following step, the resulting sinograms were reconstructed with the original CT-based μ-map in order to obtain a reference image. Likewise, reconstruction was performed using the μ-maps derived from each pseudo CT.

3.8 Discussion and conclusion

This chapter presents the first main contribution to this thesis, a novel deep learning framework for MR to CT synthesis, namely Deep Boosted Regression. The method was compared to a state-of-the-art multi-atlas propagation method, a popular deep neural network (U-Net) and a high-resolution compact network architecture (HighRes3DNet). In order to quantify the results, the Mean Absolute Error (MAE) and the Mean Squared Error (MSE) of the synthesized CT images were calculated. Only voxels within the head region were consid-

Table 3.1: Mean Absolute Error (MAE) in pCT generated with HighRes3DNet and imitation learning pCTs and corresponding MAE in pPET in the brain region only and in the whole head for all five folds.

Model	MAE pCT (in HU)	MSE pCT (in HU^2)	MAE pPET (in a.u.)
Multi-Atlas	150.96 ± 52.40	91316.86 ± 47790.80	174.68 ± 79.01
U-Net	115.12 ± 11.31	85116.06 ± 12792.78	155.39 ± 59.11
HighRes3DNet	71.76 ± 5.48	25904.41 ± 3434.87	122.92 ± 27.18
DBR	68.26 ± 3.13	20037.24 ± 1366.79	85.87 ± 22.15

ered by masking the surrounding air out. The choice in error metrics derives from its good suitability for CT synthesis and due to the quantitative nature of CT images. Furthermore, PET images were reconstructed using each pseudo CT as attenuation map. The results are demonstrated in Table 3.1.

All three deep learning-based methods show superior performance compared to the multi-atlas method (MAE: 150.96 HU $\pm$ 52.40 HU). A visual comparison between all four pseudo CT images and their corresponding MAE and MSE can be seen in Fig. 3.12. All methods are able to reconstruct realistic looking pseudo CT images with errors mainly

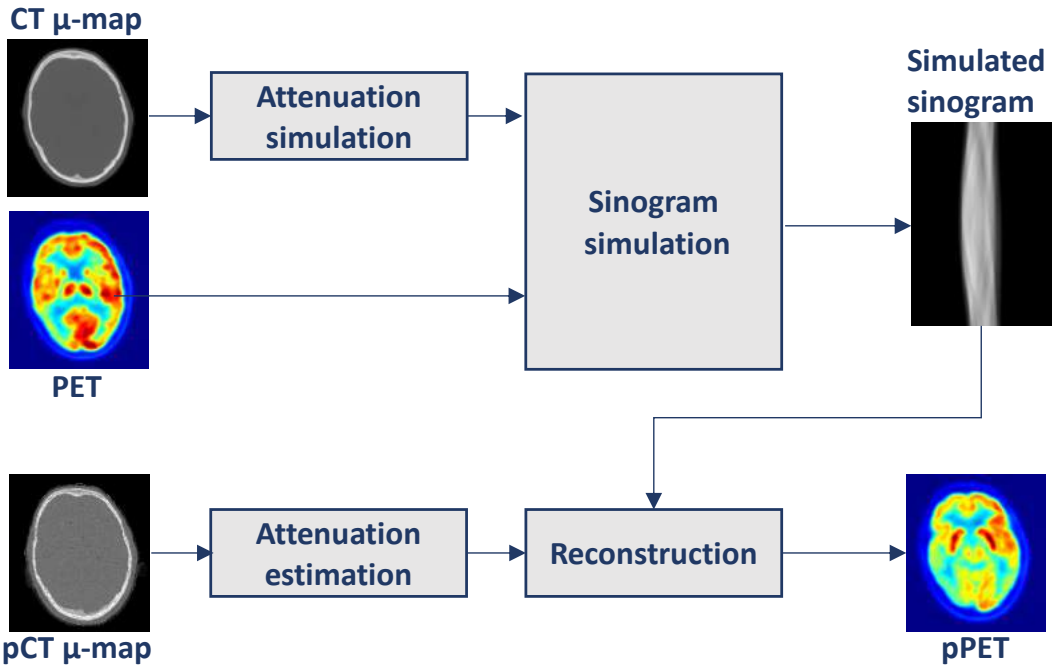

Figure 3.11: PET simulation: a PET forward projection is applied on the μ-map transformed CT to obtain attenuation factor sinograms. Similar forward projection is applied to the original PET to obtain simulated emission sinograms. Final pPETs are reconstructed from simulated emission sinograms using pCT derived attenuation maps.

within the skull region. This is to be expected as bone density cannot be definitely determined from an MR image, i.e., bone has a value of 0 in the MR image, but can have a range of values in the CT depending on how dense the bones are. The three deep learning methods are able to reconstruct the epidermoid cyst in the subject's skull whereas the multi-atlas propagation method fails to recognize this distinct anatomical feature. It is important to note that no other subject in the MR/CT database had a similar anatomical abnormality, which explains why the multi-atlas information propagation method struggles to recognize this distinct feature. As the multi-atlas propagation method is based on the registration and fusion of images within the MR/CT database, it is not possible to reconstruct any abnormal feature that does not exist within the database. This also applies to other abnormalities like tumors. On the contrary, the deep learning methods learn a spatially aware mapping function between MR and CT. When comparing the three deep learning methods, it can be seen that U-Net performs worst generating pseudo CT images with a MAE of 115.12 HU $\pm$ 11.31 HU and a MSE of 85116.06 HU $\pm$ 12792.78 HU compared to pseudo CT images synthesized with HighRes3DNet only (71.76 HU $\pm$ 5.48 HU and 25904.41 HU $\pm$ 3434.87 HU) and the proposed Deep Boosted Regression (68.26 HU $\pm$ 3.13 HU and 20037.24 $\pm$ 1366.79).

Evaluating the effect of the pseudo CT images when used as attenunation maps for

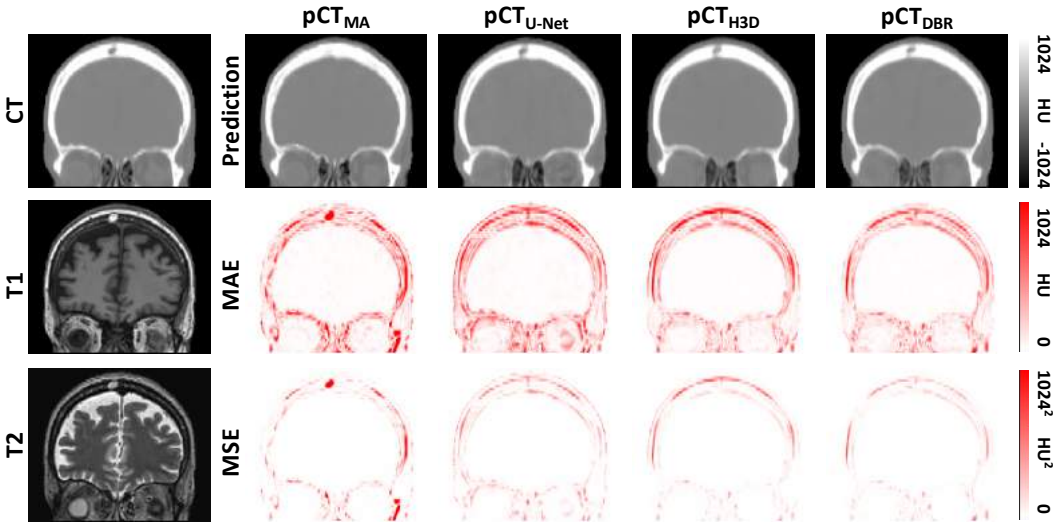

Figure 3.12: Ground truth CT and input T1- and T2-weighted MR images (first column) followed by predicted pseudo CT images with corresponding Mean Absolute Error (MAE) and Mean Squared Error (MSE) for multi-atlas propagation, U-Net, HighRes3DNet and Deep Boosted Regression.

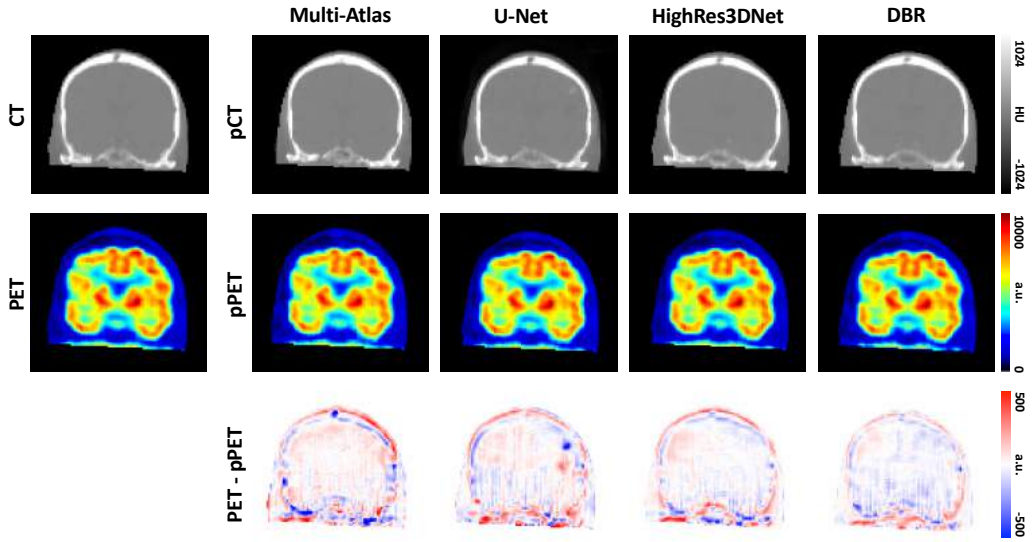

Figure 3.13: Ground truth CT and PET reconstructed with attenuation map derived from ground truth CT (first column) followed by predicted pseudo CT images with corresponding PET reconstructed with attenuation map derived from each pseudo CT and corresponding PET reconstruction error for multi-atlas propagation, U-Net, HighRes3DNet and Deep Boosted Regression.

PET reconstruction, it can be seen that the PET error can be linked to the error in the CT image. The PET reconstruction error is highest for the PET reconstructed with the pseudo CT generated by the multi-atlas propagation method (174.68 a.u. $\pm$ 79.01 a.u.) compared to the deep learning methods. However, the relation between the errors is not linear. The MAE of the pseudo CT generated with the HighRes3DNet only is more than 50% lower than the one from the pseudo CT generated with the multi-atlas propagation method. However, the PET reconstruction error of the PET image reconstructed with the pseudo CT generated with the HighRes3DNet shows an improvement of only 11% (multi-atlas propagation: 174.68 a.u. $\pm$ 79.01 a.u., HighRes3DNet: 155.39 a.u. $\pm$ 59.11 a.u.). The proposed DBR method performs best when evaluating the PET reconstruction error, achieving a MAE of 85.87 a.u. $\pm$ 22.15 a.u., which is more than 50% lower than the error in the PET reconstructed with the pseudo CT synthesiszed with the multi-atlas propagation method. Qualitative results can be seen in Fig. 3.13 and confirm the quantitative results.

To summarize, a novel CNN for MR-to-CT image translation was introduced that is

able to synthesize CT images from input MR images by gradually reducing the error using a separate boosting network. A four-fold random bootstrapped validation showed that this method outperforms state-of-the-art multi-atlas propagation and deep learning methods. The method's performance is superior in both pseudo CT synthesis and subsequent PET reconstruction. DBR is able to learn a spatially aware mapping from MR to CT images to realistically generate abnormalities present in the target image but absent in the source domain. This is evidenced in the example of an epidermoid cyst in the skull. No subject in the training dataset showed a remotely similar abnormality yet DBR reconstructed this unseen feature with great accuracy. However, Cohen et al. even showed that CNNs trained with losses such as the classical $\mathcal{L}_2$-loss are able to predict abnormalities such as tumors despite the training dataset not containing any tumor images (Cohen et al. 2018). This shows that neural networks are a powerful tool for solving the image-to-image translation task. However, the success of the training highly depends on the registration quality of the MR/CT database. Even small inaccuracies in the registration can have a great influence on the training. An idea to circumvent the requirement of paired data is to incorporate a generative adversarial loss which provides a means of learning the context between CT and MR images from unpaired data. This has potential to provide a significant advantage in terms of data availability for training due to the scarcity of accurately paired datasets, however, challenges in terms of validation emerge due to a missing ground truth. Furthermore, in their work, Cohen et al showed that distribution matching losses used in generative models can hallucinate image features, i.e., they can translate a healthy brain image into one that contains tumors. Additionally, each training patch only saw a small part of the training image, which could potentially lead to the network failing to learn sufficient contextual information. This could be circumvented by using larger training patches subject to more powerful hardware.

Chapter 4

Multimodal learning

The CNN described in chapter 3 was trained on a database of co-registered T1- and T2-weighted MR and CT images. This way, the network can use information contained in both T1- and T2-weighted MR images to approximate a mapping function to the CT image domain. However, many datasets only contain one type of MR images. Therefore, the following chapter explores the impact of the MR input modality used during training. Three networks with the exact same dataset split were trained using either T1- or T2-weighted MR images as input or both.

4.1 T1-weighted images

T1-weighted images highlight the fatty tissues of the body such as subcutaneous fat on the scalp. On the contrary, fluid filled spaces in the body (e.g. cerebrospinal fluid (CSF) in the brain) appear dark in T1-weighted MR images due to their lack of fat. Bone, air and the CSF all have low intensities in T1-weighted images, which can cause difficulties when registering T1-weighted images to CT images, because all three tissue classes differ greatly from each other in CT images. In order to explore if a CNN is able to overcome these difficulties and still learn the relation between T1-weighted MR and CT images, a DBR model was trained with T1-weighted input images. The data used in this chapter is the same dataset used in chapter 3. The network was trained on 15 co-registered T1-weighted MR and CT images. Two images were used for validation to avoid overfitting and the network performance was evaluated on three test images (equivalent to a 75% training 10% validation 15% testing split). The synthesis results with the corresponding ground truth CT and the residual between them can be seen in Fig. 4.1.

It can be seen that the network struggles to learn bone as well as the mostly homogenous soft tissue intensities inside the cranial vault. The brain structure from the T1-weighted

MR image is visible in the generated pseudo CT. The residual image confirms the visual results. The network reconstructs a pseudo CT image that continuously underestimates the bone intensities within the skull. Compared to the bone error, the error within the cranial vault appears to be low, however, they are high enough to be apparent in the pseudo CT image. Furthermore, the network does not recognize the sinus region and fails to reconstruct the small bone structures within this region.

4.2 T2-weighted images

Unlike T1-weighted images, T2-weighted images not only recognize images of fatty tissue but also water-based tissue. This means that the CSF appears hyperintense in T2-weighted images, allowing a better differentiation between high CSF intensities and low bone/air intensities. A second DBR network was trained until convergence with the same configurations and dataset split as the previous experiment to explore the ability of the network to find a mapping between T2-weighted MR images and CT images. Figure 4.2 shows the synthesis results with the corresponding ground truth CT and the residual between them.

Looking at the synthesis results of the DBR network trained with T2-weighted input images only, it can be seen that just like in the previous experiment the error is focused around the skull region. However, it appears that the error is generally lower and the network rather overestimates the density within the skull. The ventricles are visible to some degree but much less obvious than in the DBR network trained with T1-weighted images only. The largest error can be seen in the whole sinus area.

4.3 T1- and T2-weighted images

The last experiment that was performed was the original DBR network with two input channels trained on T1- and T2-weighted images. This way, the network can take advantage of both imaging contrasts and has more information available to find a mapping function be-

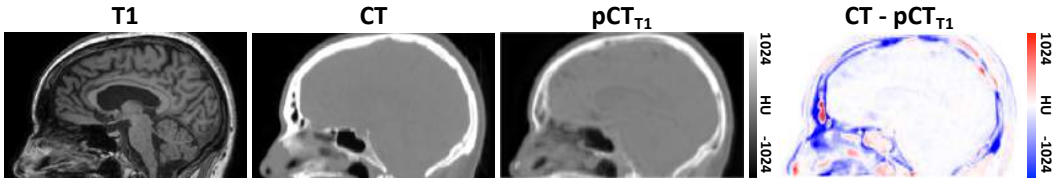

Figure 4.1: From left to right: T1-weighted MR input image, ground truth CT image, predicted pseudo CT, residual between ground truth CT and pseudo CT.

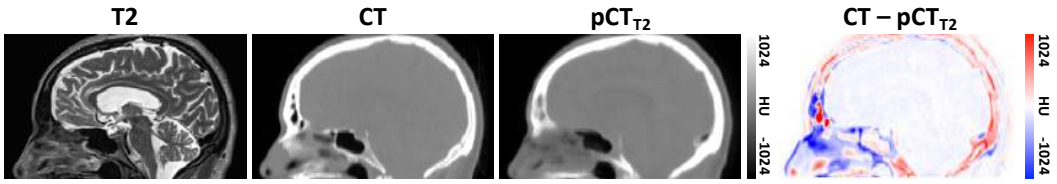

Figure 4.2: From left to right: T2-weighted MR input image, ground truth CT image, predicted pseudo CT, residual between ground truth CT and pseudo CT.

tween MR and CT. Figure 4.3 shows the synthesis results alongside the ground truth CT and the corresponding synthesis error.

The DBR network trained with T1- and T2-weighted input images shows the best results visually. The skull appears realistic and sharp. This observation is confirmed by the residual map, the error within the skull region is lowest for a network trained with multi-channel input. However, the network performance within the sinus region shows a high error similar to the previous experiments. Overall, the generated pseudo CT looks most realistic and shows the lowest residual when trained with both T1- and T2-weighted input images.

4.4 Discussion and conclusion

In order to validate the above experiments, the MAE and the MSE were calculated for each subject and averaged for all three models. Results of this validation are depicted in Table 4.1.

It can be seen that all three models perform best on the second test subject, and that for all three subjects the MAE is highest when trained on T1-weighted images only. This observation is consistent with the problems known from registering T1-weighted images to CT images. The network struggles to learn the correct CT intensities for the skull and the CSF, which have a similar proton density within the T1-weighted MR image. Although there are obvious errors in the generated pseudo CT, the network learns the spatial context of the

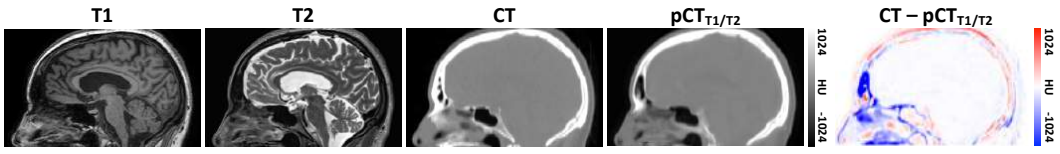

Figure 4.3: From left to right: T1-weighted MR input image, T2-weighted MR input image, ground truth CT image, predicted pseudo CT, residual between ground truth CT and pseudo CT.

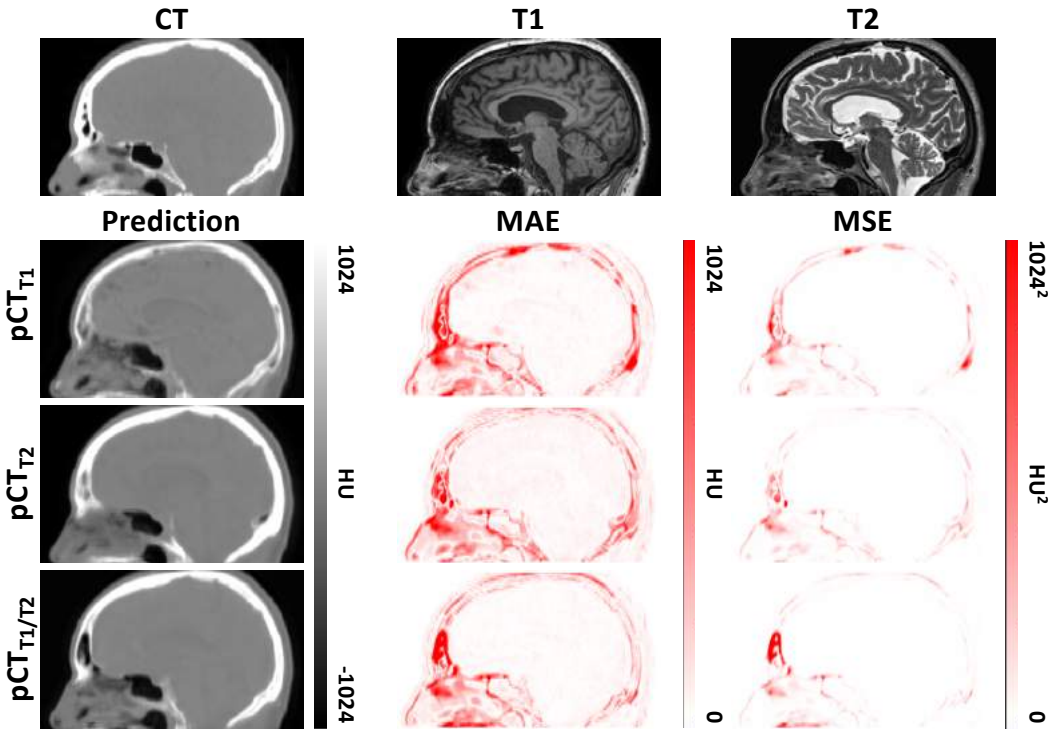

Figure 4.4: Mean Absolute Error (MAE) and Mean Squared Error (MSE) in pseudo CTs generated with HighRes3DNet trained with T1-weighted, T2-weighted and both T1- and T2-weighted MR images as network input.

images and therefore does not misclassify the CT intensities completely. This observation is confirmed when looking at the MAE image in Fig. 4.4. When training the DBR model with T2-weighted input images only, the pseudo CT improves visually within the skull region as well as in the cranial vault. Looking at the numbers in Table 4.1 it can be seen that the error is generally lower than compared to the model trained with T1-weighted images only. Lastly, the DBR model trained with a combination of T1- and T2-weighted images

Table 4.1: Mean Absolute Error (MAE) and Mean Squared Error (MSE) in pseudo CTs generated with HighRes3DNet trained with T1-weighted, T2-weighted and both T1- and T2-weighted MR images as network input.

Subject	MAE pCT (in HU)			MSE pCT (in HU^2)		
	T1-weighted	T2-weighted	T1- & T2-weighted	T1-weighted	T2-weighted	T1- & T2-weighted
1	134.82	112.11	90.83	68920.80	31423.72	51904.51
2	88.01	87.93	64.86	43209.71	25677.52	30366.76
3	111.87	94.80	77.86	35488.00	16399.72	29192.35
Average	111.57 ± 23.41	98.28 ± 12.46	77.85 ± 12.98	49206.17 ± 17504.46	24500.32 ± 7580.86	37154.54 ± 12787.34

outperforms the networks trained with single modality input for all three test subjects. The MAE within the cranial vault is minimal and the residuals in the skull region are smaller than in the other two models. The MAE of the dual modality input model has an average value of 77.85 $\pm$ 12.98 compared to the model trained with T2-weighted images with an average MAE of 98.28 $\pm$ 12.46 and the the model trained with T1-weighted images with an average MAE of 111.57 $\pm$ 23.41.

All three models show the highest error within the sinus region. This is expected as it is hard for the network to differentiate between small bones and air in this region. In both T1- and T2-weighted images the sinus region is not distinct, thus making it hard for the network to assign the correct intensity. This further explains the high MSE in the sinus regions. Due to its quadratic nature the MSE highlights the areas with the errors that contribute most to the error metric. This is particularly obvious when looking at the axial slides depicted in Fig. 4.5. It can be observed that all networks have problems reconstructing the small structures within the sinus region highlighted by the blue errors. In the model trained with T1- and T2-weighted images the errors within the sinus region have the biggest contribution towards the total error, whereas in the single modality trained models an additional error source stems from the posterior part of the skull.

Overall, it is clear that superior synthesis performance is achieved when training a DBR network with dual modality input compared to training a network with only T1- or T2-weighted input images respectively.

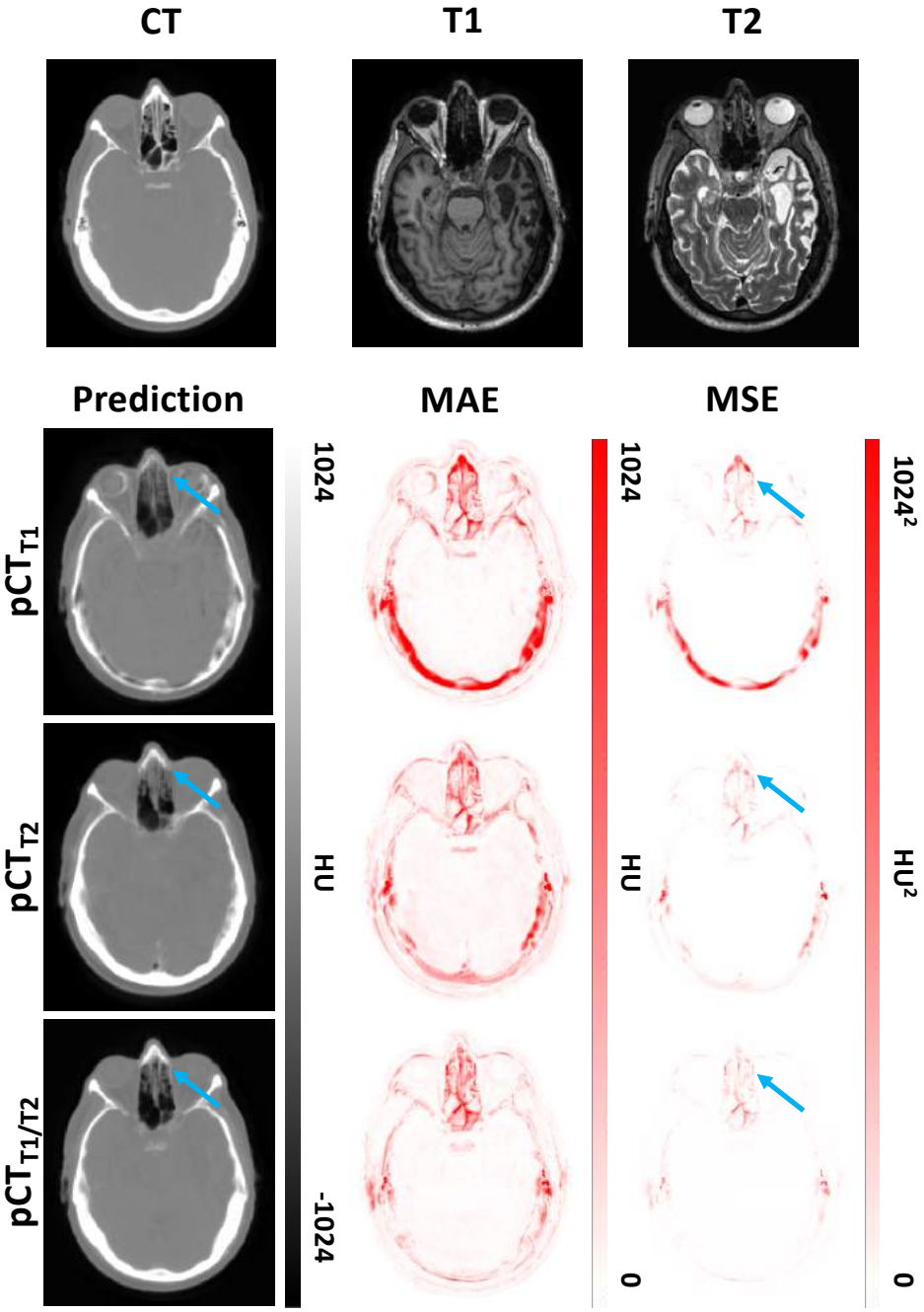

Figure 4.5: Mean Absolute Error (MAE) and Mean Squared Error (MSE) in pseudo CTs generated with HighRes3DNet trained with T1-weighted, T2-weighted and both T1- and T2-weighted MR images as network input demonstrated on axial slides. Blue errors highlight problems in sinus region.

Chapter 5

Whole-body CT synthesis

Up to this point, all work was developed on a database of head MR and CT images based on the availability of data. However, the initial aim of this project was to develop novel methods for PET/MR attenuation correction in the thoracic region. Therefore the work within this chapter focuses on the development of a neural network for pseudo CT synthesis for whole-body images. Until 2019, the problem of whole-body MR to CT synthesis, especially in 3D, has largely remained untackled.

In 2019, Dong et al. presented a method for whole-body PET/MR attenuation correction where they estimate pseudo CT images from non-attenuation corrected PET images (Dong et al. 2019). They employed a CycleGAN framework in tandem with a self-attention strategy to generate whole-body pseudo CT images. In the same year, Hwang et al. (Hwang et al. 2019) published a pseudo CT synthesis method for whole-body PET/MR attenuation correction. Their method utilizes a U-Net style neural network that takes activity and attenuation maps, estimated using the MLAA algorithm (see section 2), to learn a CT-derived μ-map. Ge et al. (Ge et al. 2019) were the first to attempt to translate full-body MR images to CT images by introducing a multi-view adversarial learning scheme that predicts 2D pseudo CT images along three axes (i.e., axial, coronal, sagittal). 3D volumes are obtained for each axis by stacking 2D slices together before an average fusion is performed to obtain one final 3D volume. The synthesis performance is then evaluated on sub-regions of the body (lungs, femur bones, spine etc). They do not, however, provide results on the full volume.

At the start of the development of the presented whole-body MR to CT image translation method in early 2019, no attempts had been made to synthesize CT images from MR images in a three-dimensional manner. It is desirable to train neural networks for medical

image analysis in 3D due to the three-dimensional nature of medical images. 3D networks are able to learn the contextual information between slices, which is of particular interest in cross-sectional imaging. 2D networks on the contrary only look at a single slice, wherefore they inherently fail to capture context from adjacent slices. The same applies for 2.5D imaging as the training is performed on 2D slices and a 3D volume is reconstructed by stacking all 2D image slices at inference time. The proposed method not only synthesizes whole-body pseudo CT images in 3D, but also estimates corresponding uncertainty maps that account for model and data uncertainty and has been published in (Kläser et al. 2020).

5.1 Data pre-processing

The dataset used for training and cross-validation consisted of 32 pairs of whole-body MR (voxel size $0.67 \times 0.67 \times 5\ mm^3$) and CT images (voxel size $1.37 \times 1.37 \times 3.27\ mm^3$). CT images were acquired on a GE Discovery 710 PET/CT scanner (140 kVp, 32mA) and T1- and T2-weighted images were acquired on a Siemens Biograph mMR PET/MR immediately after. T1-images were acquired using a two-point three-dimensional volumetric interpolated breath-hold examination (VIBE) Dixon sequence (3.0 T; TE/TR, 1.23 ms/4.02 ms; flip angle 10°) and T2-weighted images were acquired using an echo-planar fast spin echo sequence (HASTE) (3.0 T; TE/TR, 107 ms/700 ms; flip angle 90°). Whole-body MR images were acquired in four/five stages consisting of 40 slices each. The standard clinical acquisition protocol for whole body PET/MR imaging suggests to perform an axial scan from skull vertex to mid thigh over four to five stations, depending on the height of the patient, which are subsequently composed. Images were of size $640 \times 500 \times 160$ or $640 \times 500 \times 200$ voxels depending on whether four or five stages had to be acquired. Acquisition times varied across patients, depending on their height, between 60-90 minutes. Both T1-weighted VIBE and T2-weighted HASTE images were acquired at breath-hold. All patients within this dataset were scanned as part of a research study comparing the diagnostic performance of 18F-FDG PET/CT to PET/MR in adult patients with suspected or proven cancers (over 15 different cancer types) leading to a wide range of pathologies. Patients were both biological sexes, male and female, and in the age range between 20 and 87 years with a mean age of 58 years. In order to generate a continuous MR image, where the four/five stages could no longer be distinguished, the MR images were pre-processed in two steps. Firstly, the bias-field within each MR image was corrected. Secondly, the four distinct stages were fused using a percentile-based intensity harmonization approach. All images were then resampled to

CT resolution before the co-registered database was built. MR and CT images were aligned using first a rigid registration algorithm followed by a very-low-degree-of-freedom non-rigid deformation. A second non-linear registration was performed, using a cubic B-spline with normalized mutual information to correct for soft tissue shift. In clinical practice, patients are required to keep their arms up in the CT in order to reduce the radiation dose, while arms are kept close to the body in MR acquisition due to the limited bore size. This deformation is so large, that registration algorithms struggle to compensate for the bone and tissue shift between the two images. A common practice that is also regularly used in radiotherapy treatment planning is to mask out the arms in both images and perform the registration on the thorax region only. Registration quality was carefully assessed manually by multiple specialists for each subject. The success of the supervised learning methods highly depends on the registration quality of the MR/CT database. Even small inaccuracies in the registration can influence the training and subsequently lead to underestimation of the attenuation map, wherefore careful registration quality assessment is essential. With perfect registration being particularly challenging in the whole body, it is of utmost importance to account for uncertainty in the model to account for such inaccuracies. Both CT and MR images were rescaled to be between 0 and 1 for increased training stability. The data were split into 70% training, 10% validation and 20% testing data for all methods.

5.2 Direct CT synthesis

As a first step, two direct synthesis networks were trained on the whole-body data: U-Net and HighRes3DNet. Figure 5.1 and Fig. 5.2 show example synthesis results of one subject and the corresponding residuals for both methods respectively. Additionally, a comparison of all methods can be seen in Fig. 5.8. The pseudo CT synthesiszed with U-Net reconstructs bone to a certain extent but is characterized by a large degree of blurriness most evident in the femurs and the ribs. The model further fails to reconstruct any bone structures in the shoulders as well as any structures within the lung. The corresponding residuals show that the highest source of error originates from underestimating the CT intensities of the lung and the shoulder bones, whereas soft tissue is generally slightly overestimated.

The pseudo CT reconstructed with HighRes3DNet looks visually similar to the results synthesized with U-Net. Rib bones are more prevalent, but appear slightly blurrier. When looking at the residuals it can be seen that the network struggles to reconstruct bones. Areas around misclassified bones also show a high error similar to a halo. The network seems

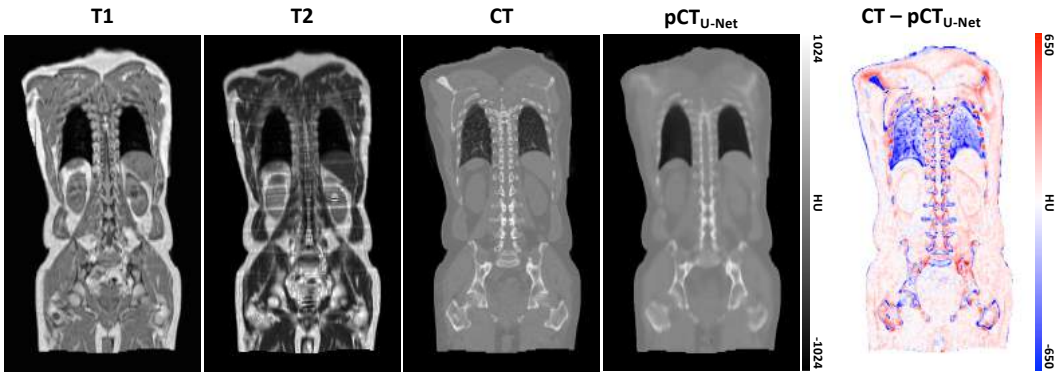

Figure 5.1: Qualitative results on whole-body data for U-Net. From left to right: T1- and T2-weighted MR input images, ground truth CT, synthesized pseudo CT and corresponding residual.

overly confident and reconstructs bones with high confidence in the wrong place. It implies that the network tries to compensate for the understimation of the bones by assigning higher intensities to tissues surrounding the understimated bone, thus the halo effect. The network fails to reconstruct structures within the lung, however, the error within the lungs is generally lower than in the pseudo CT reconstructed with U-Net. Overall intensities within the pseudo CT reconstructed with HighRes3DNet are slightly overestimated.

5.3 DBR for whole-body CT synthesis

As a second step, the method proposed in chapter 3, Deep Boosted Regression, was applied to the whole-body synthesis problem. Figure 5.3 shows an example of the synthesis result

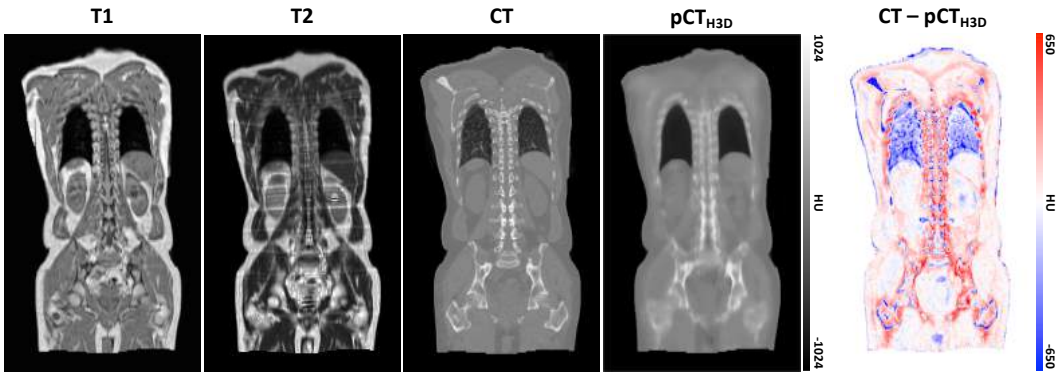

Figure 5.2: Qualitative results on whole-body data for HighRes3DNet. From left to right: T1- and T2-weighted MR input images, ground truth CT, synthesized pseudo CT and corresponding residual.

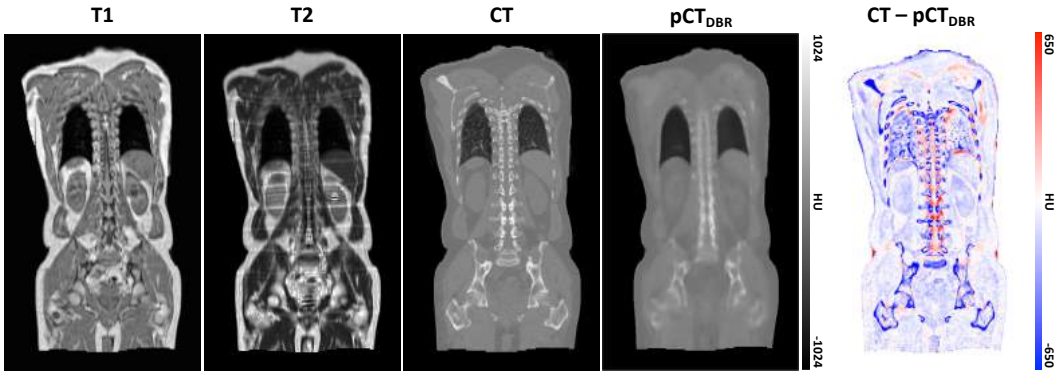

Figure 5.3: Qualitative results on whole-body data for Deep Boosted Regression. From left to right: T1- and T2-weighted MR input images, ground truth CT, synthesized pseudo CT and corresponding residual.

and the corresponding residuals. The synthesized pseudo CT looks visually blurrier than the two direct synthesis models (U-Net and HighRes3DNet). Similar to the results from U-Net, the model fails to reconstruct bones, in particular in the shoulder region. The residuals show that intensities are generally underestimated and the highest source of error arises from the lack of reconstructed detail within the bones.

5.4 Multi-scale network for whole-body CT synthesis

The main challenge with whole-body data is its size, and the fact that a large field of view is necessary to make accurate predictions. Common networks, such as a U-Net, can only store patches of size $160 \times 160 \times 160$ due to 32GB VRAM GPU memory limitations. To tackle this issue, an end-to-end multi-scale convolutional neural network is proposed that takes input patches from full-body MR images at three resolution levels as inputs to synthesize high resolution, realistic CT patches. The network also incorporates explicit heteroscedastic uncertainty modelling by casting the task likelihood probabilistically, and epistemic uncertainty estimation via traditional Monte Carlo dropout. A patch-based training approach is employed whereby at each resolution level of the network a combination of downsampling and cropping operations results in patches of similar size but at different resolutions, spanning varied fields of view. Three independent instances of HighRes3DNet are trained simultaneously, thus not sharing weights, taking patches of each resolution as input each resulting in a feature map with different resolution. Lower level feature maps are concatenated to those at the next level of resolution until the full resolution level, where these con-

catenated feature maps are passed through two branches of $1 \times 1 \times N$ convolutional layers resulting in a synthesized CT patch and the the corresponding voxel-wise heteroscedastic uncertainty. Note, that the $1 \times 1 \times N$ convolutional layers are indeed part of the original HighRes3DNet architecture. Once low, middle and high resolution feature maps are concatenated, the $1 \times 1 \times N$ convolutional layer decodes the concatenated feature maps to a corresponding CT image patch. This is illustrated in Fig. 5.4. Similarly to Kamnitsas et al. (Kamnitsas et al. 2017), the hypothesis is that such a design allows the network to simultaneously benefit from the fine details afforded by the highest resolution patch and the increased spatial context provided by the higher field of view prominent in the low resolution patches. They proposed a multi-scale, 3D CNN architecture for brain lesion segmentation that consists of eleven layers. The network consists of two pathes that learn image features at different scales that are concatenated before they are passed through a fully connected layer and an additional classification layer in order to create a lesion segmentation. Thus, during optimization, the network tries to minimize the Cross Entropy between predicted and actual segmentation. The proposed MultiRes network consists of an additional path, wherefore image features are learned at three resolutions. Furthermore, the proposed MultiRes network incorporates an additional level of deep supervision for each resolution that Kamnitsas et al. (Kamnitsas et al. 2017) misses. Finally, the second branch of $1 \times 1 \times N$ convolutional layers allows to compensate for heteroscedastic uncertainty in the model.

5.4.1 Modelling heteroscedastic uncertainty

Previous works on MR to CT synthesis have shown that residuals are not homogeneously spread throughout the image, rather, they are largely concentrated around bone/organ/tissue boundaries. As such, a heteroscedastic uncertainty model is most suitable for this task, where data-dependent, or intrinsic uncertainty is assumed to be variable. Heteroscedastic uncertainty is also called data-inherent uncertainty and describes the internal randomness of a given phenomenon. For medical images, such as MR and CT images, this data-inherent uncertainty corresponds to intrinsic noise in the observational data, e.g. noise from detectors, and cannot be compensated by acquiring more data. When modelling heteroscedastic uncertainty, the first step is to model the pseudo CT synthesis task likelihood as a normal distribution with mean $f^{W}(x)$, which is the model output corresponding to the input $\mathbf{x}$, parameterized by weights $\mathbf{W}$, and pixel-wise standard deviation $\sigma^{W}(x)$. The data intrinsic noise can then be described as a predictive distribution such that:

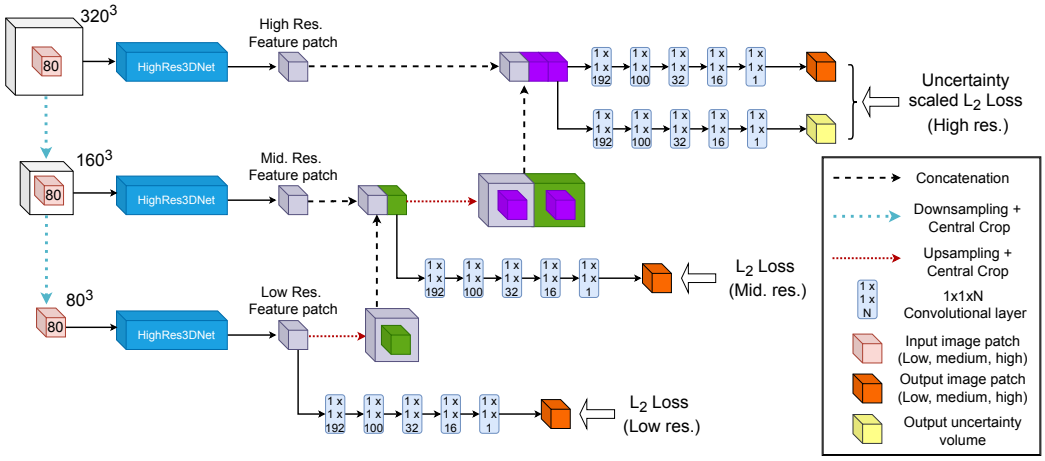

Figure 5.4: Proposed MultiRes network architecture. T1- and T2-weighted MR patches of size 320^3 are fed into the HighRes3DNet architecture at various levels of resolution and field of view. Lower level feature maps are concatenated to those at the next level until the full resolution level, where these concatenated feature maps are passed through two branches consisting of a series of $1 \times 1 \times N$ convolutional layers: one resulting in a synthesized CT patch and the other to the corresponding voxel-wise heteroscedastic uncertainty.

$$p(y|f^W(x)) = \mathcal{N}(f^W(x), \sigma^W(x)) \tag{5.1}$$

The loss function is subsequently derived by calculating the negative log of the likelihood:

$$\begin{aligned} \mathcal{L}(y,x;\mathbf{W}) &= -\log p(y|f^W(x)) \\ &\approx \frac{1}{2\sigma^W(x)^2}\left(y - f^W(x)\right)^2 + \log \sigma^W(x) \\ &= \frac{1}{2\sigma^W(x)}\mathcal{L}_2(y, f^W(x)) + \log \sigma^W(x) \end{aligned} \tag{5.2}$$

In those regions where the observed $\mathcal{L}_2$ error remains high, the uncertainty should compensate for the error and also increase the uncertainty. The second term in the loss prevents the collapse to the trivial solution of assigning a large uncertainty everywhere.

5.4.2 Modelling epistemic uncertainty

Model uncertainty, also called epistemic uncertainty, arises when a model is not trained optimally, mostly caused by a lack of training data. This means that an increased amount

of data can reduce model uncertainty. In the scenario, where an infinite amount of data is available, the model uncertainty can be explained away to zero. However, many deep learning scenrios, lack training data altogether, wherefore it is important to model epistemic uncertainty. Test-time dropout has been established as the go-to method for estimating model uncertainty, a Bayesian approximation at inference. By employing dropout during training and testing it is possible to sample from a distribution of sub-nets that in the regime of data scarcity will provide varying predictions. This variability captures the uncertainty present in the network's parameters, allowing for a pixel-wise estimation by quantifying the variance across these samples.

Traditional dropout has been introduced in 2014 by Srivastava et al. as a means to avoid overfitting during training (Srivastava et al. 2014). Overfitting a model in the context of CNNs can be understood as an overly confident network, that means the network performs well on a particular training set, but fails to produce reliable predictions on unseen data. The network "memorizes" training samples instead of estimating a generalizable model. Therefore, it is desirable to have a large amount of training data, however, this is often not the case in real life scenarios. Traditional dropout can help to avoid overfitting by randomly dropping out network weights during training time. A visual depiction of the idea behind dropout is presented in Fig. 5.5. Each node is associated with a dropout probability p that determines how likely it is that a weight is voided. In each training iteration, a new set of weights is sampled, thus preventing the network from memorizing training samples.

During test time, however, no weights are voided resulting in a deterministic prediction. This means, the model will always give the exact same label or value for one input.

In *Monte Carlo Sampling* (MCS) or *Monte Carlo (MC) Dropout* (Gal & Ghahramani 2016), weights are also dropped during testing time such that the output is no longer deterministic. This means that a given datapoint can result in different output values when the model is applied multiple times. MC dropout can be seen as drawing samples from a probabilistic pseudo CT distribution.

Here, channel dropout was chosen over the traditional neuron dropout. Channel dropout has indeed been shown to be better for convolutional layers where channels fully encode image features while neurons do not encode individually such meaningful information (Hou & Wang 2019). Dropout samples at inference time are acquired by performing N stochastic forward passes over the network, equivalent to sampling from the posterior

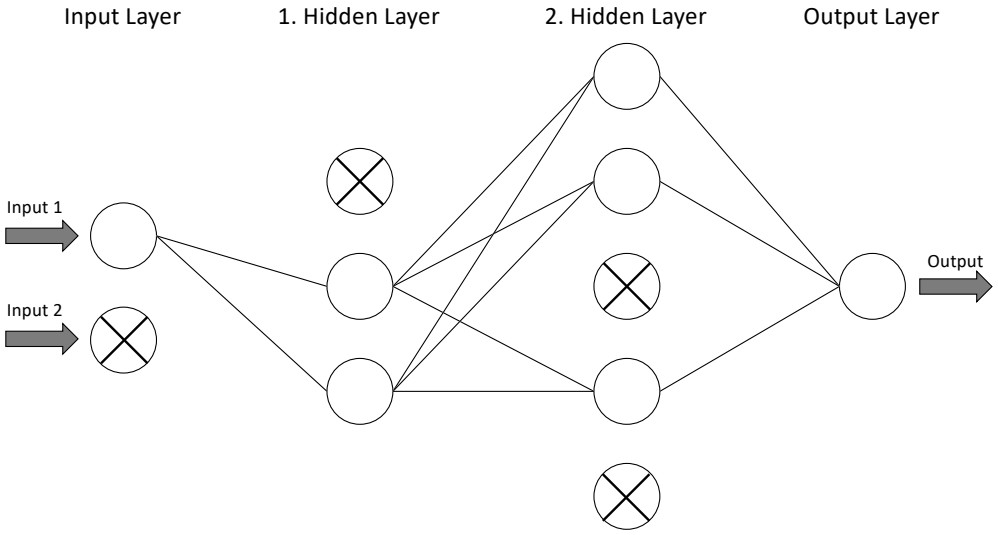

Figure 5.5: Dropout. During training time, random weights of the network are voided in order to avoid overfitting. At each training iteration, a different set of weights is dropped out. Crossed nodes have been dropped out, thus set to 0.

over the weights. A measure of uncertainty can be obtained by calculating the variance over these samples on a pixel-wise basis.

5.4.3 Implementation details

The multi-scale network consists of three residual networks, each taking in a $80 \times 80 \times 80$ MR image patch with different resolutions and fields of view. In order of high, medium, and low resolution, the MR patches are obtained by taking an initial high resolution $320 \times 320 \times 320$ patch and cropping the central $80 \times 80 \times 80$ region (high), downsampling the initial patch by a factor of two and taking the central $80 \times 80 \times 80$ patch (medium), and finally downsampling the initial patch by a factor of four to obtain a $80 \times 80 \times 80$ patch (low).

Starting from the lowest resolution sub-net, the output of size $80 \times 80 \times 80$ is upsampled by a factor of two and centrally cropped. This patch is concatenated with the output of the medium resolution sub-net. This concatenated patch of size $80 \times 80 \times 80 \times 2$ is then upsampled by a factor of two and centrally cropped, before being concatenated to the output of the high-resolution sub-net. These series of upsamplings and crops ensure that the final

outputs contain patches with the same field of view prior to the final set of four 3D convolutions of kernel size 1, which produce the CT patch. Image patches of size $80 \times 80 \times 80$ were used due to a limited GPU memory budget. Larger input images could potentially increase training performance because the network can learn the contextual information better, however, this is dependent on available hardware.

Heteroscedastic variance is modelled by the addition of a series of four $1 \times 1 \times 1$ convolutional layers following the concatenation of the combined low-medium scale output to the high scale output, architecturally identical to the convolutional layers for the synthesis branch. Channel dropout probability (i.e., the probability to keep any one channel in a kernel) was set to 0.5, both during training and testing, and N=20 forward passes were carried out for each experiment. Parameters were chosen with regards to recommendations in original literature. The batch size was set to one, Adam was used as the optimizer and networks were trained until convergence, where this was defined as a sub 5% loss change over a period of 5000 iterations.

5.4.4 Qualitative results

As a first initial experiment the proposed MultiRes network was trained without both epistemic and heteroscedastic uncertainty in order to see if the information provided by multiple levels of resolution can help to achieve superior synthesis results. A qualitative example is shown in Fig. 5.6. The pseudo CT looks reasonably sharp and shows good bone delineation. Compared to the results generated with HighRes3DNet, U-Net and DBR, the pseudo CT synthesized with the MultiRes network seems much sharper in the ribs and the femurs. Furthermore, it is the only model that attempts to reconstruct the shoulder bones. Looking at the residuals, it can be seen that, similar to the other models, the highest errors stem from bones and lungs. The network appears to be overly confident in its bone prediction thus predicting bone incorrectly. The error in the remaining body parts (organs, muscle tissue, etc.) is, however, very low. Compared to the other synthesis results the overall error is a combination of over- and underestimation of CT intensities.

As a second step, the proposed uncertainty-aware MultiRes$_{unc}$ network was trained in order to ascertain if the additional uncertainty information can improve the network's performance. Results are demonstrated in Fig. 5.7. The pseudo CT looks blurry, similar to the results generated with U-Net, HighRes3DNet and DBR. Bone structures can be seen, but are less sharp than in the pseudo CTs generated with MultiRes without uncertainty.

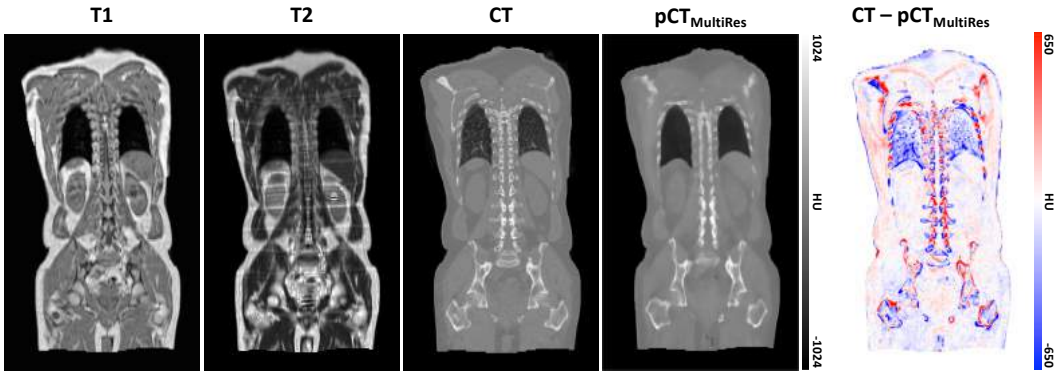

Figure 5.6: Qualitative results on whole-body data for proposed MultiRes network without uncertainty. From left to right: T1- and T2-weighted MR input images, ground truth CT, synthesized pseudo CT and corresponding residual.

Just like U-Net, HighRes3DNet and DBR, the uncertainty-aware MultiRes$_{unc}$ model fails to reconstruct the shoulder bones completely. However, when looking at the corresponding residuals, it can be seen that the network performs better than U-Net and DBR. The overall error is low and evenly distributed between under- and overestimation. The highest source of error arises from the lack of information reconstructed in the bone leading to a high underestimation in those regions. Interestingly, the proposed uncertainty-aware MultiRes$_{unc}$ network shows the best synthesis performance in the lungs when compared to all other models.

5.5 Discussion and conclusion

This chapter presents the second main contribution to this thesis, a novel uncertainty-aware multi-resolution deep learning framework for MR to CT synthesis especially developed for the use of whole-body data. The network can further be trained without uncertainty, if desired. The method was compared to two deep networks that directly synthesize pseudo CTs, U-Net and HighRes3DNet, and the boosting network presented in chapter 3. In order to quantify the results the Mean Absolute Error (MAE) and the Mean Squared Error (MSE) of the synthesized CT images were calculated only within the body by masking the surrounding air out. The results are shown in Table 5.1.

Quantitative results show that the two direct neural networks (U-Net and HighRes3DNet) have the highest MAE of 92.89 $\pm$ 13.30 HU and 89.05 $\pm$ 8.77 HU respectively. However, HighRes3DNet has fewer trainable variables and is therefore more efficient. A visual depiction of the MAE is shown in Fig. 5.8. For U-Net the main source of error arises

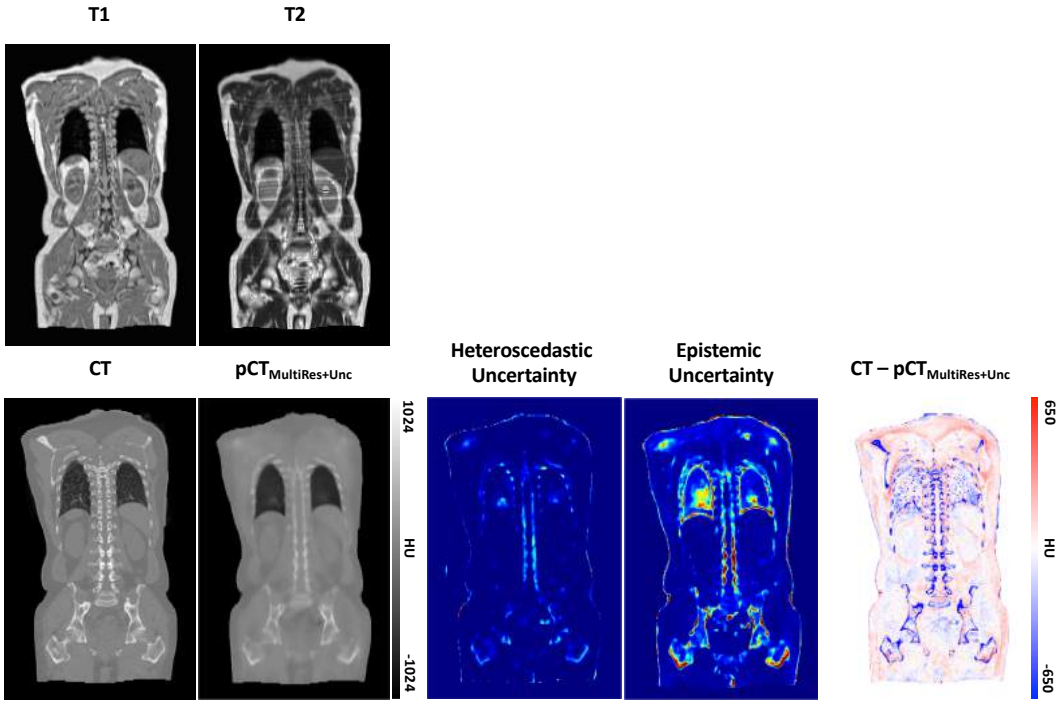

Figure 5.7: Qualitative results on whole-body data for proposed MultiRes network including uncertainty estimation. From left to right: T1- and T2-weighted MR input images, ground truth CT, synthesized pseudo CT with heteroscedastic and epistemic uncertainty, and corresponding residual.

Table 5.1: MAE and MSE across all experiments including number of trainable variables. Bolded entries denotes best model (p-value < 0.05).

Experiments	Model parameters	MAE (HU)	MSE (HU^2)
3D U-Net	14.49M	92.89 $\pm$ 13.30	37358.07 $\pm$ 11266.56
HighRes3DNet	0.81M	89.05 $\pm$ 8.77	23346.09 $\pm$ 3828.22
DBR	1.62M	77.58 $\pm$ 3.20	19026.56 $\pm$ 2779.69
MultiRes	2.54M	**72.87** $\pm$ **2.33**	18532.23 $\pm$ 1538.41
MultiRes$_{unc}$	2.61M	73.90 $\pm$ 6.24	**16007.56** $\pm$ **2164.76**

from bones and the lungs while soft tissue and organs have a low error. On the contrary, HighRes3DNet shows a lower MAE in the lungs but a higher error in tissues surrounding bone. The MSE, also depicted in Fig. 5.8, highlights the regions of the pseudo CTs that have the highest overall contribution to the error.

Both methods struggle to reconstruct bone properly, likely because of the lack of spatial

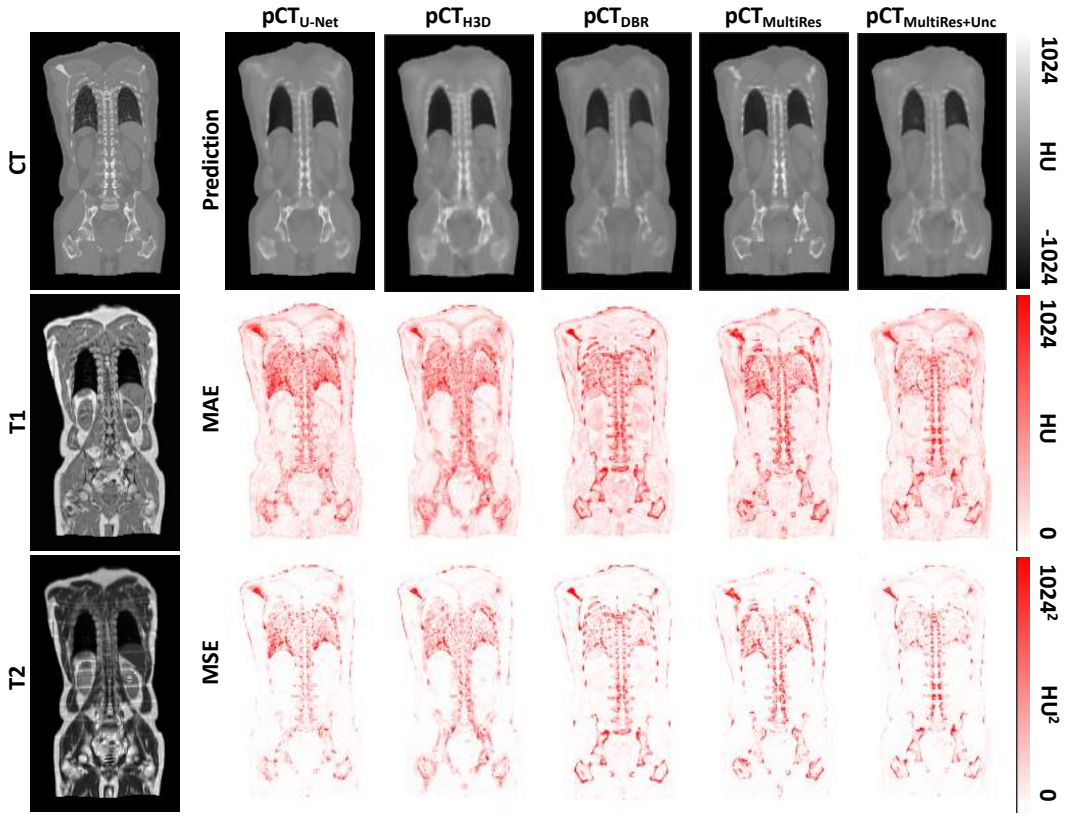

Figure 5.8: Ground truth CT and input T1- and T2-weighted MR images (first column) followed by predicted pseudo CT images with corresponding Mean Absolute Error (MAE) and Mean Squared Error (MSE) for U-Net, HighRes3DNet, Deep Boosted Regression, proposed MultiRes without uncertainty and proposed MultiRes$_{unc}$ including uncertainty estimation.

context necessary to reconstruct small (relatively) cohesive structures such as vertebrae. The Deep Boosted Regression approach proposed in chapter 3 shows a superior performance compared to the direct synthesis networks with a MAE of 77.58 $\pm$ 3.20 HU. Bone structures show an even higher error than U-Net and HighRes3DNet, prevalent in the MSE map in Fig. 5.8, however, the recursive boosting nature of this network minimizes the overall error of the reconstructed pseudo CT. Just like U-Net and HighRes3DNet, DBR suffers from the limited field of view of the patch-based training approach making it difficult for the network to recognize small distinct structures.

The proposed MultiRes model exhibits the greatest bone fidelity: the individual vertebrae are clearer, with intensities more in line with what would be expected for such tissues, and the femurs boast more well-defined borders. This model further shows the lowest MAE of 72.87 $\pm$ 2.33 HU. The proposed MultiRes$_{unc}$ model leads to blurrier results than the

simpler proposed MultiRes model without uncertainty, likely due to the inclusion of the loss uncertainty term and limited network capacity. Both MultiRes models have a higher number of trainable variables and are therefore slightly less efficient than HighRes3DNet by itself, however, the afforded performance increase compensates for this. The MultiRes$_{unc}$ model demonstrates similar bone reconstruction as U-Net, HighRes3DNet and DBR, however, the overall MAE of 73.90 $\pm$ 6.24 HU is significantly lower. It is interesting to note, that the MSE of the proposed MultiRes$_{unc}$ model is lower than the MSE of its uncertainty unaware counterpart. This shows that the majority of residuals in the pseudo CT generated by MultiRes$_{unc}$ are in a lower range than for MultiRes without uncertainty and therefore when squared do not contribute as much to the MSE. This is likely due to the fact that the network is less confident in bone regions, whereas the models that do not compensate for uncertainty are overly confident and predict high bone intensities in the wrong place resulting in a high error. A possible cause for this are mis-registration errors between the three imaging modalities. Although T1- and T2-weighted MR images were aquired in the same imaging session, natural deformation of organs can occur, such as lung deformation due to breathing, bowel movement, continuous heart beat and others. These natural deformations are even more prevalent in the CT images as the time between imaging sessions is larger since the patient needs to be transfered into a different imaging suite. Furthermore, the patient's position within an MR scanner generally differs from the position of the patient in a CT scanner. Due to the limited bore size and the long acquisition times, patients are scanned with their arms down in MR scanners, whereas they are scanned with arms up in CT scanners most of the time in order to reduce the amount of radiation that the body is exposed to. The differences in bone position and soft tissue deformation are so large that it is difficult to align MR and CT images perfectly. In some cases registration even fails completely. Therefore, although the majority of the arms were masked out for this dataset, the deformation of the thorax due to the different arm position is difficult to compensate.

The uncertainty-aware MultiRes$_{unc}$ model shows high uncertainty in those regions and therefore has low confidence in reconstructing bone. The uncertainty-aware MultiRes$_{unc}$ model is particularly less confident in the shoulder regions where the highest reconstruction errors are expected due to the differing acquisition protocols of MR and CT imaging. Furthermore, small registration inaccuracies, such as air bubbles in the bowel, are captured by the uncertainty-aware model that can cause visually worse synthesis results. While bone

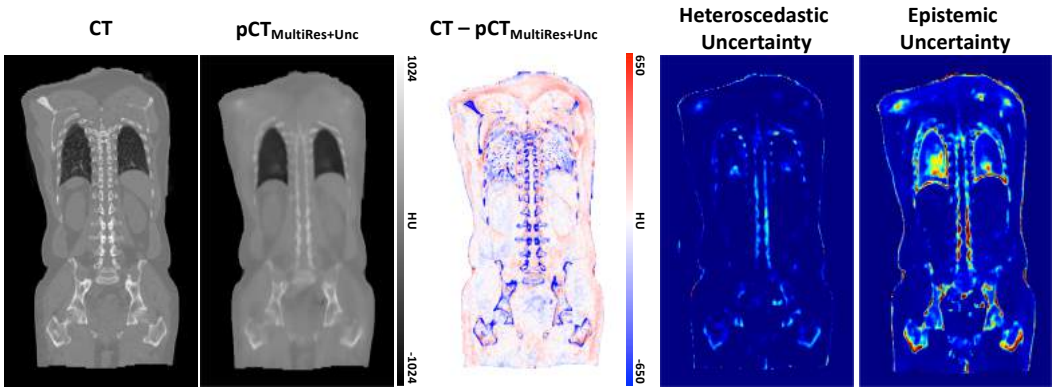

Figure 5.9: From left to right: CT ground truth, pseudo CT prediction of MultiRes$_{unc}$, corresponding residuals, heteroscedastic uncertainty and epistemic uncertainty. Both uncertainties correlate with the absolute error map.

boundaries are generally very distinct in CT images, they are not clear in MR images. In fact, manually segmenting bone in MR images leads to high inter-observer errors, wherefore a high uncertainty is expected to be assigned by the network in such regions. This can further be observed when looking at the uncertainty of the MultiRes$_{unc}$ model shown in Fig. 5.9. There is a strong correlation between uncertainty and residuals, which suggests that the model appropriately assigns a higher uncertainty to those regions that are difficult to predict. Both epistemic and heteroscedastic uncertainties exhibit large values around structure borders, as expected. The borders between tissues are not sharp and there is, therefore, some ambiguity in these regions, which is mirrored by the corresponding overlapping error in the residuals. In theory, an increased amount of data should diminish the epistemic uncertainty by providing the network with a greater number of samples from which to learn the correspondence between MR and CT in these areas. The aforementioned blurriness, however, could result in some inconsistency in the synthesis process, which would still be captured by the heteroscedastic uncertainty. The benefits afforded to MultiRes$_{unc}$ for being uncertainty-aware can further be seen in Fig. 5.10. The joint histograms are constructed by calculating the error rate, taken as the difference between the ground truth CT and pseudo CT averaged across N=20 dropout samples, at different levels of both epistemic and heteroscedastic uncertainty (standard deviations per voxel) and taking the base 10 log. The red line describes the average error rate at each level of uncertainty. A strong correlation between uncertainty and error rate can be observed, suggesting that the model appropriately assigns a higher uncertainty to those regions that are difficult to predict.

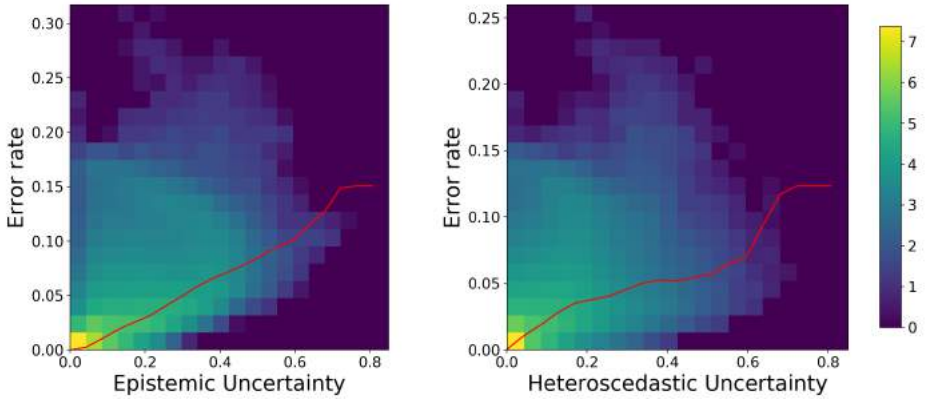

Figure 5.10: Joint histogram of prediction uncertainty and error rate for proposed MultiRes$_{unc}$ network: epistemic (left), heteroscedastic (right). The average error rate at different uncertainty levels is shown by the red line. Error rate tends to increase with increasing uncertainty, showing that the network correlates uncertainty to regions of error.

It is interesting to note that the degree of uncertainty is particularly high in the vicinity of air pockets. Unlike corporeal structures, it is expected that these pockets are subject to more deformation between the MR and CT scanning sessions, resulting in a lack of correspondence between the acquisitions in these regions. This results in the network attempting to synthesize a morphologically different pocket to what is observed in the MR, resulting in a high degree of uncertainty. The same applies to misregistered areas of the body as well as natural soft tissue deformations.

To summarize, the contributions to this chapter are two-fold: MultiRes, a novel learning scheme for multi-resolution MR to CT synthesis of the full body, and MultiRes$_{unc}$, an extension to this model that incorporates uncertainty as a safety measure and to account for intrinsic data noise. A significantly superior performance (p-value < 0.05) of MultiRes and MultiRes$_{unc}$ can be observed by comparing the proposed methods to single-resolution CNNs, U-Net and HighRes3DNet as well as the previously proposed Deep Boosted Regression. Furthermore, the importance of modelling uncertainty is demonstrated, showing that MultiRes$_{unc}$ is able to identify regions where the MR to CT translation is most difficult.

In a data-scarce environment, it becomes especially important to quantify uncertainty as networks are unlikely to have sufficient evidence for full convergence. After all, accurately aligning CT and MR images is inevitable to validate the voxel-wise performance of any image synthesis algorithm until other appropriate methods have been developed that allow validating on non-registered data. Despite the slightly decreased performance of

MultiRes$_{unc}$ compared to MultiRes, both from a quantitative and qualitative standpoint, the additional clinical insight introduced by modelling uncertainty is a valuable asset. Furthermore, while the model does not reconstruct bone-based structures as well as its uncertainty agnostic counterpart, it still outperforms all three baseline models. Therefore, the slight decrease in performance of the uncertainty-aware model is insignificant compared to the important additional information provided by the uncertainty.

Chapter 6

End-to-end optimization

6.1 Limitations of CT-based losses

Up to this point, the main objective has been to minimize the error between the predicted pseudo CT and the corresponding ground truth CT, which is equivalent to minimizing the $\mathcal{L}_2$-loss. This objective was justified by the fact that in current clinical practice the gold standard for PET/MR attenuation correction is an additional CT acquisition that can be linearly rescaled to an attenuation map used in PET reconstruction. $\mathcal{L}_2$-losses are a sensible loss metric when the optimal pseudo CT for PET reconstruction is the one that is in terms of intensity the closest to the target ground truth CT. However, $\mathcal{L}_2$-losses do not recognize that the main aim of CT synthesis, when used for PET/MR attenuation correction, is to generate a synthetic CT that, when used as attenuation map for PET reconstruction, makes the reconstructed PET as close as possible to the gold standard PET reconstructed with the true CT. Furthermore, the risk-minimizing nature of $\mathcal{L}_2$-losses (e.g., the sinus region, which is dark in T1-weighted MR images, can be mapped to air or to bone but not to any value in between (Cardoso et al. 2015)) ignores small local differences between the predicted pseudo CT and the ground truth CT, which can significantly impact the reconstructed PET image. An illustration of this downstream impact in PET reconstruction can be seen in Fig. 6.1.

At the time of developing the following CT synthesis approach, all CT synthesis methods found in the literature concentrate on minimizing the residuals of the predicted pseudo CT. However, pseudo CT synthesis only represents an interim stage when intended to correct for photon attenuation in PET/MR and thus creating an additional space for likely introduced errors. The aim of the proposed method is to directly minimize the PET residuals. This is achieved by introducing a novel MR to CT synthesis framework that is composed of two separate CNNs. The first CNN synthesizes multiple valid CT predictions

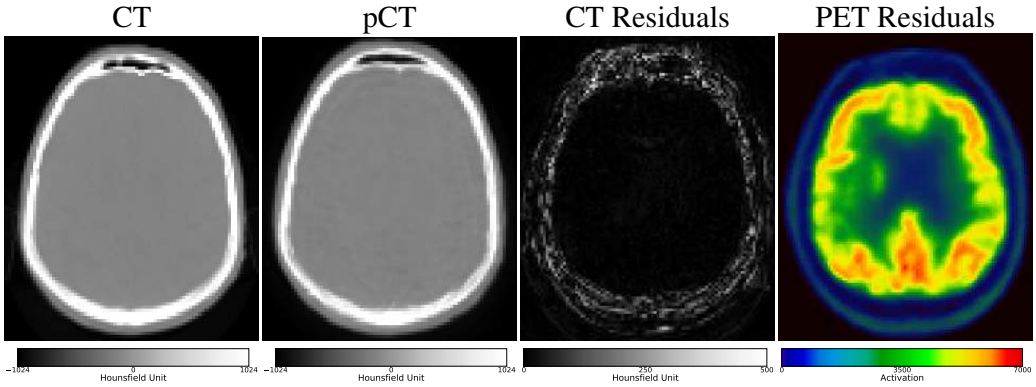

Figure 6.1: a) Ground truth CT, b) predicted pseudo CT, c) absolute error between ground truth and pseudo CT, and d) absolute error between PETs reconstructed using the ground truth CT and pseudo CT for attenuation correction. Small and very localized differences in the CT (c) can lead to large errors in the PET image (d). Therefore, CNNs should optimize for PET residuals (d) and not for CT residuals (c) when used for PET attenuation correction.

using multi-hypothesis learning instead of a single pseudo CT only (Rupprecht et al. 2017). Multi-hypothesis learning deals with the fact that uncertainty is inherent in the image synthesis task. Instead of assuming only a single prediction, multi-hypothesis learning allows for multiple plausible predictions, similar to Monte Carlo dropout (see chapter 5). Once the network predicted multiple plausible CT representations, an oracle determines the predictor that generates the most correct pseudo CT and only updates the weights with regards to the winning mode. An oracle can be seen as a ”black box” that can produce a solution for any computational problem (here i.e. finding the best pseudo CT predictor). This enables the first CNN to specialize in predicting pseudo CTs with distinct features (e.g., skull thickness, bone density). A second CNN then uses imitation learning to predict the residuals between ground truth PETs and PETs reconstructed with each valid pseudo CT. Imitation learning tries to mimic human behavior of any given task (here i.e. imitate the PET reconstruction process). In this setting, the second CNN acts as a metric that predicts the pseudo PET residuals. By minimizing this metric loss, the network learns to synthesize pseudo CT images that will ultimately result in pseudo PETs with lower residuals.

6.2 Sampling for multiple realizations

The proposed method assumes that there are multiple valid mappings from MR to CT as there is no definite function that can describe the correlation between MR and CT images (air and bone have the same intensity in MR images, but opposite intensities in CT). Distinct

anatomical features in the CT such as skull thickness and bone density cannot definitively be determined from an MR image as proton density simply does not give any conclusion on these factors. Therefore, two different sampling strategies have been implemented in the following work: *Monte Carlo Sampling* and *Multi Hypothesis Sampling*.

6.2.1 Monte Carlo sampling

The first method that was implemeted in order to synthesize multiple realistic realizations of pseudo CTs is called MC dropout (see chapter 5.4.2). In this work, three forward passes were performed to generate three different realizations of plausible pseudo CTs. It is not trivial to model for epistemic and heteroscedastic uncertainty in the framework proposed in this chapter, wherefore MC dropout was exclusively used to generate multiple pseudo CT realizations from a probabilistic pseudo CT distribution.

6.2.2 Multi-hypothesis sampling

The second method that was implemeted in order to synthesize multiple realistic realizations of pseudo CTs is called *Multi-Hypothesis Sampling*. Multi-hypothesis sampling was introduced by Rupprecht et al. in 2017 as an interim step to estimate uncertainty (Rupprecht et al. 2017). During training, multiple pseudo CT realizations are generated by looping M times through the last $1 \times 1 \times 1$ convolutional layer of the network, thus resulting in a different set of weights for each of the M network outputs. An oracle then determines the predictor that generates the most correct pseudo CT (lowest $\mathcal{L}_2$-loss) and only updates the weights of the network with regards to the winning mode. This enables the network to specialize in predicting pseudo CTs with distinct features such as skull thickness or bone density.

Mathematically speaking, the proposed image synthesis approach aims to find a mapping function f_ϕ between two image domains $\mathcal{X}$ and $\mathcal{Y}$, where a set of input MR images $x \in \mathcal{X}$ and a set of output CT images $y \in \mathcal{Y}$ is given

$$f_\phi : \mathcal{X} \rightarrow \mathcal{Y} \quad \text{with} \quad \phi \in \mathbb{R}^M. \tag{6.1}$$

In a supervised learning scenario with a set of N paired training tuples (x_i, y_i), $i = 1, ..., N$, the network tries to find the predictor f_ϕ that minimizes the error

$$\frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f_\phi(x_i), y_i). \tag{6.2}$$

$\mathcal{L}$ can be any desired loss. Just as in the previous presented methods, the classical $\mathcal{L}_2$-loss was chosen. In the proposed multi-hypothesis scenario, the network provides multiple predictions of valid pseudo CT realizations:

$$f_\phi^j(x) \in [f_\phi^1(x), ..., f_\phi^M(x)] \quad \text{with} \quad M \in \mathbb{N}. \tag{6.3}$$

As in the original work for multi-hypothesis learning, only the loss of the best predictor $f_\phi^j(x)$ will be used during training following a Winner-Takes-All (WTA) strategy, i.e.,

$$\mathcal{L}(f_\phi(x_i), y_i) = min_{j\in[1,M]}\mathcal{L}(f_\phi^j(x_i), y_i). \tag{6.4}$$

This way the network learns M modes to generate pseudo CT images, where each mode specializes on specific features. M was set to 3 as it has been shown to be large enough to capture the task uncertainty (Rupprecht et al. 2017).

6.2.3 Comparison

In order to find a more suitable sampling scheme between the above presented methods, the use of either method was evaluated: MC dropout versus multi-hypothesis learning. The results are demonstrated in Fig. 6.2. The intensities of pseudo PETs reconstructed with a μ-map from pseudo CTs generated with MC dropout show an artificially low variance, whereas the intensities of pseudo PETs reconstructed with the pseudo CTs synthesized with the multi-hypothesis model provide a wider distribution. To investigate the accuracy of the predictions, the Z-score of the ground truth PET with regards to each sampling scheme was calculated to demonstrate the relationship between the mean data distribution and the ground truth PET. Fig. 6.2-Right presents the per pixel Z-score defined as

$$\frac{\text{PET}-\mu(\text{pPET}^M)}{\sigma(\text{pPET}^M)}, \tag{6.5}$$

where $\mu(\text{pPET}^M)$ and $\sigma(\text{pPET}^M)$ are the per-pixel average and per pixel variance over M pseudo PET samples respectively. Results show that a significantly lower Z-score can be found in the brain region for the multi-hypothesis model in comparison to when MC dropout is used. This means that the multi-hypothesis-based PET uncertainty encompasses the true PET intensity more often than the competing MC dropout method.

Consequently, the development of the proposed imitation learning approach exclusively includes multi-hypothesis sampling in order to generate multiple pseudo CT realiza-

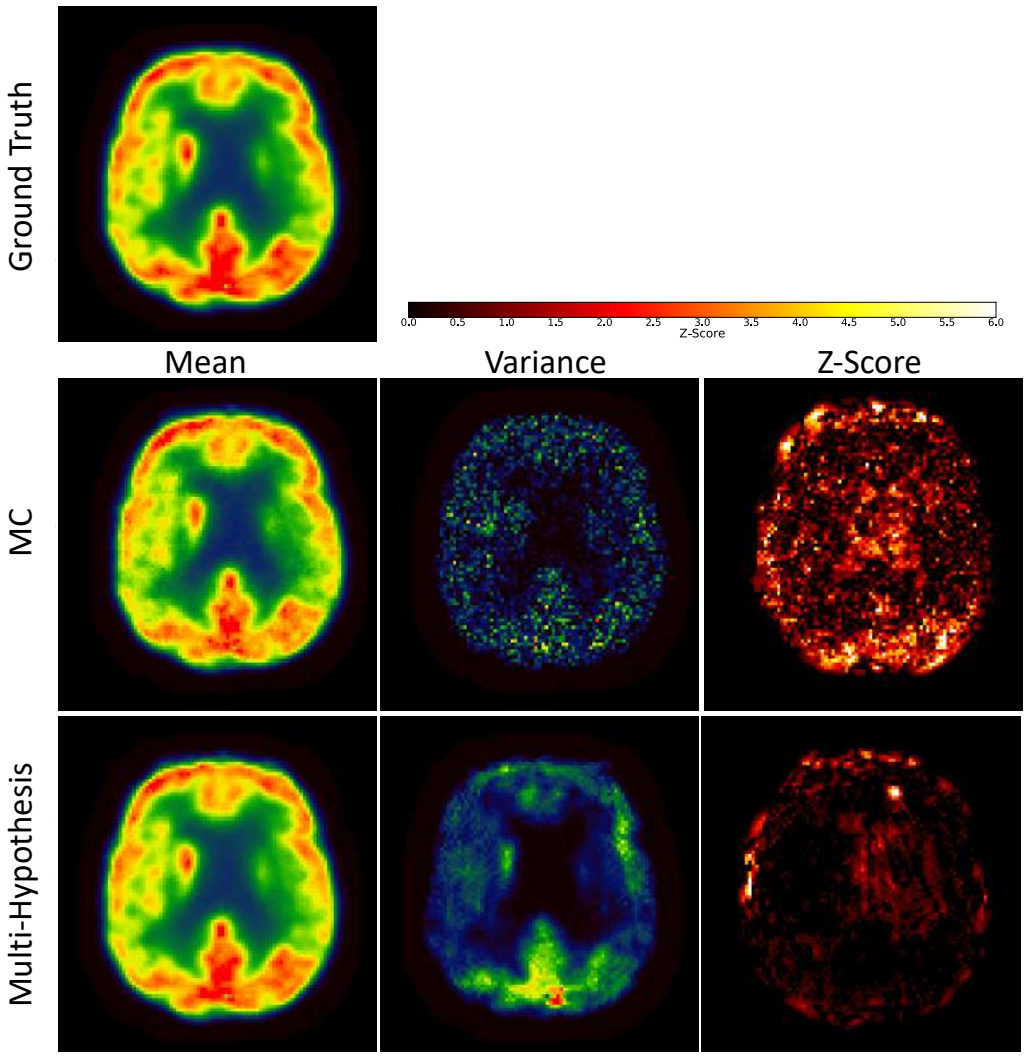

Figure 6.2: PET values (first column), variance (middle column) and Z-score (right column) of ground truth PET (top row) compared to pseudo PET values reconstructed with pseudo CTs from Monte Carlo (MC) dropout sampling (middle row) and pseudo CTs from multi-hypothesis sampling (bottom row). The multi-hypothesis model captures true PET values better than the MC dropout method.

tions.

6.3 Imitation learning for CT synthesis

Following the hypothesis that the $\mathcal{L}_2$-loss is not optimal as a loss metric for pseudo CT synthesis when used to correct for attenuation in PET/MR because of its risk minimizing nature (e.g., the sinus region which is dark in T1-weighted MR images can be mapped to air or to bone but not to any value in between), the proposed network consists of a second CNN that aims to minimize subsequent PET residuals. This network approximates the function

$$g_\psi : \mathcal{Y}, \tilde{\mathcal{Y}} \rightarrow \mathcal{Z} \quad \text{with} \quad \psi \in \mathbb{R}^n, \tag{6.6}$$

by taking ground truth CTs (y_i) and pseudo CTs ($f_\phi^j(x_i) \in \tilde{\mathcal{Y}}$) as inputs. Here, $\mathcal{Z}$ is a set of error maps between the ground truth PET and the pseudo PET that was reconstructed (as in section 3.7) with each of the M pseudo CT realizations as μ-maps. In other words, the second CNN learns to predict the PET reconstruction error from an input CT/pCT pair, thus imitating, or approximating, the PET reconstruction process. This imitation CNN is trained by minimizing the $\mathcal{L}_2$-loss between the true PET uptake error z and the predicted error $\tilde{z}$, i.e.,

$$\mathcal{L}_2 = ||z - \tilde{z}||_2. \tag{6.7}$$

Lastly, this second CNN is used as a new loss function for the first CNN and minimizes the RMSE, as it provides an approximate and differentiable estimate of the PET residual loss. Thus, the loss minimized by the first CNN is defined as

$$\begin{aligned} \mathcal{L}(f_\phi(x_i), y_i, z_i) = &\min_{j \in [1,M]} \mathcal{L}(f_\phi^j(x_i), y_i) \\ &+ \min_{j \in [1,M]} [g_\psi(f_\phi^j(x_i), y_i), z_i]. \end{aligned} \tag{6.8}$$

6.4 Proposed network architecture

The proposed network architecture is presented in the following three sub-sections. Training of the proposed imitation learning framework is performed in three distinct phases.

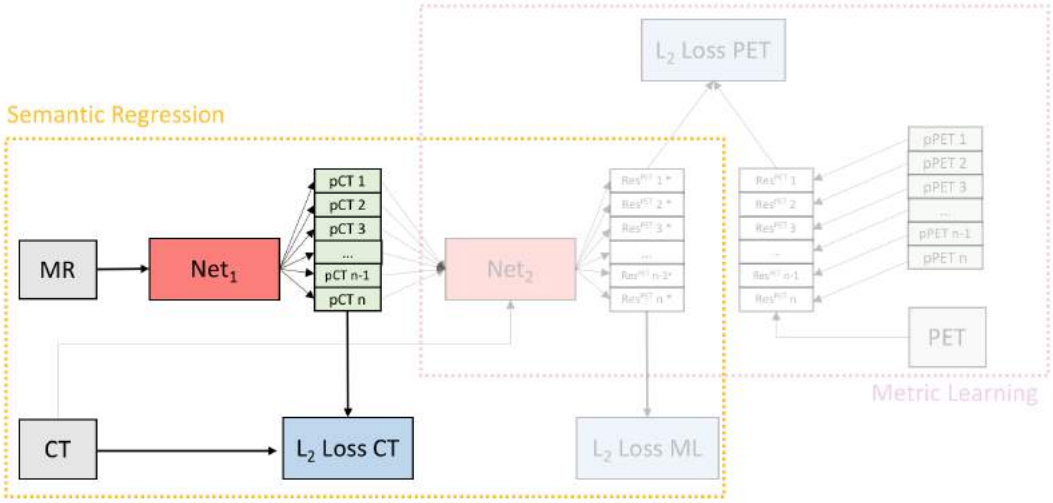

Figure 6.3: Yellow solid box: semantic regression. A first CNN (Net_1), here HighRes3DNet, with MR images as inputs predicts multiple valid pseudo CT realizations by minimizing a combination of the $\mathcal{L}_2$-loss between true CT and pseudo CT ($\mathcal{L}_2$-loss CT).

6.4.1 First training stage

In the first stage, a neural network, here HighRes3DNet, is trained with multiple hypothesis outputs minimizing an $\mathcal{L}_2$-WTA loss in order to generate different pseudo CT realizations (Fig. 6.3 yellow solid box). The first stage results in multiple realisations of pseudo CT images that act as an input to the second stage. As previously seen, multi-hypothesis learning was chosen to generate multiple realisations of pseudo CT images as it captures the true PET intensity more often than the competing MC dropout method.

6.4.2 Second training stage

In the second training stage, the weights of the first network (yellow box) are frozen and a second neural network, again a HighRes3DNet, (Fig. 6.4 purple dashed box) is trained individually. This second network takes the pseudo CT images generated in the first stage as input images and learns to predict the residual between the PET reconstructed with the true CT-derived μ-map and the pseudo PET that was reconstructed using the μ-map derived from each pseudo CT to correct for attenuation. This way the network learns the mapping between the pseudo CT residual and subsequent pseudo PET reconstruction error. Note, that all pseudo PET images are reconstructed before the training process starts, such that a paired CT/PET residual database is available for training.

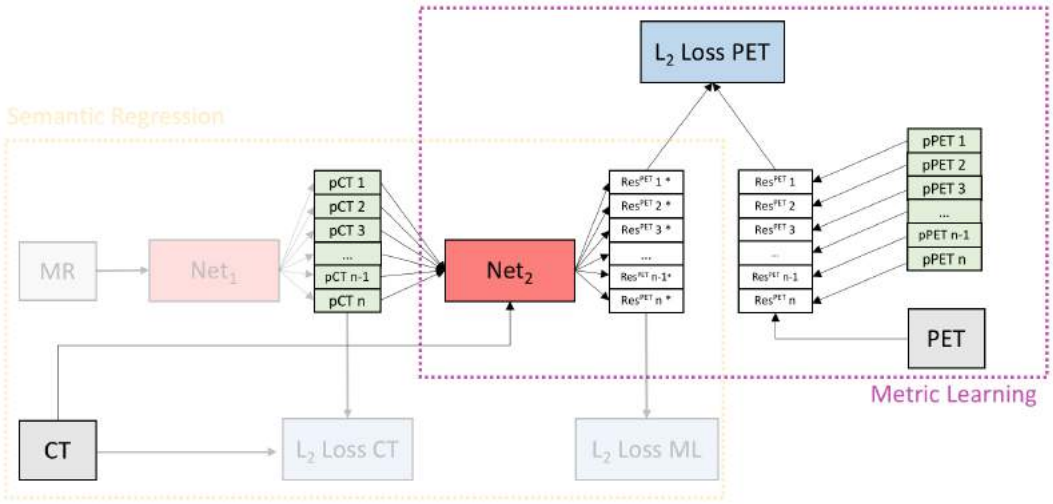

Figure 6.4: Purple dashed box: imitation network. A second CNN (Net_2), HighRes3DNet, with pseudo CTs that were generated in stage one and corresponding CTs as input predicts the residuals between PET reconstructed with true CT-derived μ-map and pseudo PET reconstructed with pseudo CT as μ-map by minimizing $\mathcal{L}_2$-loss PET. Thus, this network imitates the PET reconstruction process.

6.4.3 Third training stage

The third stage is a combination of stage one and two (Fig. 6.5. In the final stage the first network is retrained with a combination of the CT $\mathcal{L}_2$-loss from stage one, and the proposed metric loss from stage two in equal proportions. All weights in the second network are frozen such that the entire network can be used as a loss function. This way the proposed metric loss can be seen as a function that describes the PET reconstruction error. Thus, the combined loss allows the network to minimize both the CT residual and the pseudo PET reconstruction error at the same time resulting in a pseudo CT that when used for PET reconstruction will generate a PET image with minimal error.

6.4.4 Implementation details

During the training stage subvolumes of size $56 \times 56 \times 56$ pixels were randomly sampled from the input data due to a limited GPU memory budget. Those patches were augmented by randomly rotating each of the three orthogonal planes on the fly by an angle in the interval of [-10° ,10°]. Further augmentations on the MR data included random scaling by a factor between 0.9 and 1.1, random bias field augmentation of all three planes and random noise in a range between 10 SNR and 25 SNR. The data were split into 70% training, 10% validation and 20% testing data. All training phases were performed on a Titan V GPU with Adam optimizer. In the first training stage a model was trained for 50k iterations with

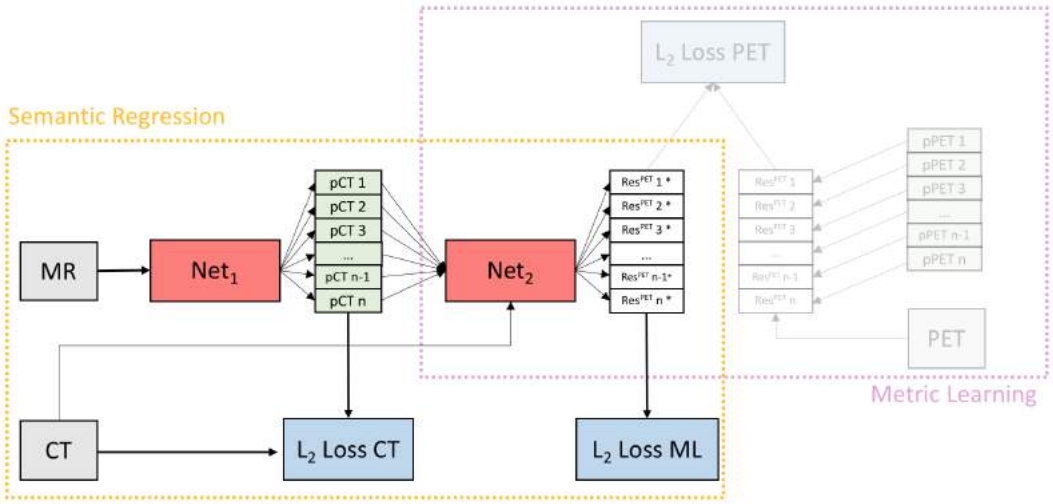

Figure 6.5: Final training stage: the first network is retrained with a combination of the CT $\mathcal{L}_2$-loss from stage one, and the proposed metric loss from stage two in equal proportions. The combined loss allows the network to minimize both the CT residual and the pseudo PET reconstruction error at the same time

a learning rate of 0.001. The network of the second training stage learning to minimize the pseudo PET reconstruction error was trained for 500k iterations with a learning rate of 0.001. During the final training stage a complete model was trained for 100k iterations minimizing a combination of the proposed losses with a learning rate of 0.001 before decreasing the learning rate by a factor of 10 and resuming training until convergence, where convergence is defined as a sub 5% change in the loss value over a period of 5000 iterations. While the number of pseudo CT images generated by the multi-hypothesis network can be chosen arbitrarily, here it was set to 3 as it has been shown to be large enough to capture the task uncertainty (Rupprecht et al. 2017).

6.5 Data pre-processing

For each subject in the training database, MRs and CTs were affinely aligned using a symmetric approach (Modat et al. 2014) based on Ourselin et al. (Ourselin et al. 2001) followed by a fully affine registration to compensate for possible gradient drift in the MR images. In the following step a very low degree of freedom non-rigid deformation was performed in order to compensate for different neck positioning before implementing a second non-linear registration, using a cubic B-spline with normalized mutual information (Modat et al. 2010). For the purpose of this work, the data were resampled to the original Siemens Biograph mMR PET resolution of $344 \times 344 \times 127$ voxels with a voxel size of approximately $2 \times 2 \times 2$ mm^3. Both CT and MR images were rescaled to between 0 and 1 for increased

training stability. Additionally, two masks were extracted for evaluation purposes, a head mask from the CT and a brain mask from the T1-weighted MR image. The head mask was generated by thresholding the CT at -500 HU thus excluding the background from the performance metric analysis. The additional brain mask was extracted from the T1-weighted MR image to exploit the radionuclide uptake in the brain region only. The data used in this chapter was the same as the data used in chapter 3 and 4. Three PETs were reconstructed with each of the multi-hypothesis pseudo CTs (here denoted as pseudo PET or pPET) in order to train the imitation CNN, resulting in a total of 60 pCT/pPET pairs.

6.6 Validation and results

Following the results of the comparison experiment between different sampling schemes (see section 6.2.3), a fully 3D model was trained and a five-fold cross-validation was performed. Qualitative results are presented in Fig. 6.6. The first column shows the true CT image (top), a pseudo CT synthesized with the HighRes3DNet chosen as baseline method (middle) and a pseudo CT synthesized using the proposed imitation learning (bottom). Next to the CTs (2nd column) the corresponding residuals between pseudo CT and true CT are illustrated. In the third column the ground truth PET (top), baseline pseudo PET (middle) and the imitation learning pseudo PET (bottom) are shown followed by the resulting pseudo PET residuals in the last column. In order to quantify the results, MAE was chosen as performance metric of the pseudo CTs only in the number of voxels in a region of interest (V), here head and brain only region. The method was not compared to U-Net as it has been shown in chapter 3.8 that HighRes3DNet achieves superior synthesis performance.

The advantages of the proposed imitation learning model were evaluated on the remaining 20% of the dataset hold out for testing (see Table 6.1). The five-fold cross-validation therefore ensures that the proposed method is trained five times on a different subset of training data and each model is tested on different test images. This way, each image in the dataset has been tested as part of the cross-validation study. The proposed method leads to a lower MAE on the CT (79.04 HU $\pm$ 3.57 HU) compared to the simple feed forward model (92.77 HU $\pm$ 8.57 HU), the MAE in the resulting pseudo PET is significantly lower (paired t-test, $p < 10^{-4}$) for the proposed method (137.61 a.u. $\pm$ 33.28 a.u. for brain region; 123.14 a.u. $\pm$ 18.38 a.u. for whole head) when compared to the baseline model (197.88 a.u. $\pm$ 69.53 a.u. for brain region; 166.36 a.u. $\pm$ 46.88 a.u. for whole head). The different results for the HighRes3DNet baseline compared to the results in chapter 3

Table 6.1: Mean Absolute Error (MAE) in pseudo CT generated with HighRes3DNet and imitation learning pseudo CTs and corresponding MAE in pseudo PET in the brain region only and in the whole head for all five folds.

Fold	MAE CT (in HU)		MAE PET brain (in a.u.)		MAE PET head (in a.u.)	
	HighRes3DNet	Imitation learning	HighRes3DNet	Imitation learning	HighRes3DNet	Imitation learning
1	103.93 ± 14.46	79.97 ± 10.19	306.98 ± 30.03	147.26 ± 26.89	240.55 ± 32.21	134.86 ± 22.34
2	99.70 ± 15.65	82.32 ± 7.91	214.63 ± 55.41	180.85 ± 69.98	180.06 ± 47.68	142.27 ± 45.69
3	89.17 ± 10.49	81.44 ± 7.49	191.21 ± 70.28	109.91 ± 18.67	122.31 ± 25.13	98.20 ± 11.62
4	86.71 ± 10.99	78.17 ± 0.22	140.31 ± 42.63	98.71 ± 12.31	154.38 ± 41.11	109.96 ± 18.70
5	84.33 ± 7.19	73.30 ± 2.26	136.27 ± 21.29	151.32 ± 26.32	134.48 ± 20.54	130.39 ± 21.43
Average	92.77 ± 8.57	79.04 ± 3.57	197.88 ± 69.53	137.61 ± 33.28	166.36 ± 46.88	123.14 ± 18.38

are potentially caused by overfitting due to the limited amount of data available.

6.7 Validating on independent head CT dataset

In order to validate the previously trained fully 3D model on a completely independent dataset, the performance of the proposed method was compared against ground truth data of 23 subjects. The method was then compared to the chosen baseline method (HighRes3DNet) and a non deep learning method, namely multi-atlas propagation that is routinely used in clinical practice and clinical trial settings. The quantitative validation was performed in two steps:

1. Pseudo CTs were synthesized from all 23 subject's MR images using the proposed method, the baseline method and the multi-atlas propagation approach. All generated pseudo CTs were then compared to the subject's ground truth CT to validate the accuracy of the synthesis.

2. Pseudo PET images were reconstructed following the simulation described in section 3.7 using μ-maps generated with pseudo CTs from proposed, baseline and multi-atlas method. All pseudo PETs were then compared to the ground truth PET that was reconstructed using the μ-map extracted from the original CT in order to validate the accuracy of the PET attenuation correction.

6.7.1 Data pre-processing

The independent validation dataset consisted of 23 subjects that were scanned on a GE Discovery 710 PET/CT scanner providing CT images (voxel size $1.367 \times 1.367 \times 3.27$ mm^3, 140 kVp, 10mA) and reconstructed ^{18}F-FDG PET images ($1.0 \times 1.0 \times 3.27$ mm^3). The 23

Table 6.2: Mean Absolute Error (MAE) in pCT generated with HighRes3DNet, multi-atlas propagation and imitation learning pCTs and corresponding MAE in pPET in the brain and head region only on independent dataset.

MAE CT (in HU)			MAE PET brain (in a.u.)			MAE PET head (in a.u.)		
Baseline	Multi-Atlas	Imitation learning	Baseline	Multi-Atlas	Imitation learning	Baseline	Multi-Atlas	Imitation learning
172.12 ± 19.61	153.40 ± 18.68	110.98 ± 19.22	642.54 ± 117.75	290.95 ± 65.46	190.05 ± 49.23	561.62 ± 87.45	289.44 ± 89.34	204.04 ± 49.03

subjects were then scanned on a Siemens Biograph mMR PET/MR immediately after. T1-weighted images were acquired using a three-dimensional magnetization-prepared rapid gradient-echo (MP RAGE) sequence (Brant-Zawadzki et al. 1992) (3.0 T; TE/TR/TI,2.63 ms/1700 ms/900 ms; flip angle 9°; voxel size $1.1 \times 1.1 \times 1.1$ mm^3). Three-dimensional isotropic T2-weighted images were acquired with a fast/turbo spin-echo sequence (SPACE) (3.0 T; TE/TR, 383 ms/2700 ms; flip angle 120°; voxel siz $1.3 \times 1.3 \times 1.3$ mm^3). Both CT and MR images were rescaled to be between 0 and 1 for increased training stability. This dataset was acquired at the end of the project, thus differs slightly from the original training dataset.

6.7.2 Imitation learning

Figure 6.6 shows the ground truth CT and pseudo CTs synthesized with the proposed imitation learning and the baseline model and the corresponding residuals as well as predicted pseudo PET images and pseudo PET residuals.

The results of the independent validation are shown in Table 6.2. The MAE over all 23 subjects in the CT for the proposed method is 110.98 HU ± 19.22 HU compared to the baseline 172.12 HU ± 19.61 HU and multi-atlas propagation method 153.40 HU ± 18.68 HU. Subsequently, the average MAE of all reconstructed PET images within the brain for the proposed method is 3.4 times lower than the MAE of the baseline (190.05 a.u. ± 49.23 a.u. compared to 642.54 a.u. ± 117.75 a.u.) in the brain region and 2.7 times lower in the whole head region (204.74 a.u. ± 49.03 a.u. compared to 561.62 a.u. ± 87.45 a.u.). Further, the proposed imitation learning method achieves an approximately 1.5 times lower average MAE of all reconstructed PET images in both the head and the brain region compared to PET images reconstructed with the pseudo CT generated with the multi-atlas propagation method (290.95 HU ± 65.46 HU in brain region, 289.44 HU ± 89.34 HU in head region).

Example images of T1-, T2-weighted, CT, pseudo CT synthesized with baseline

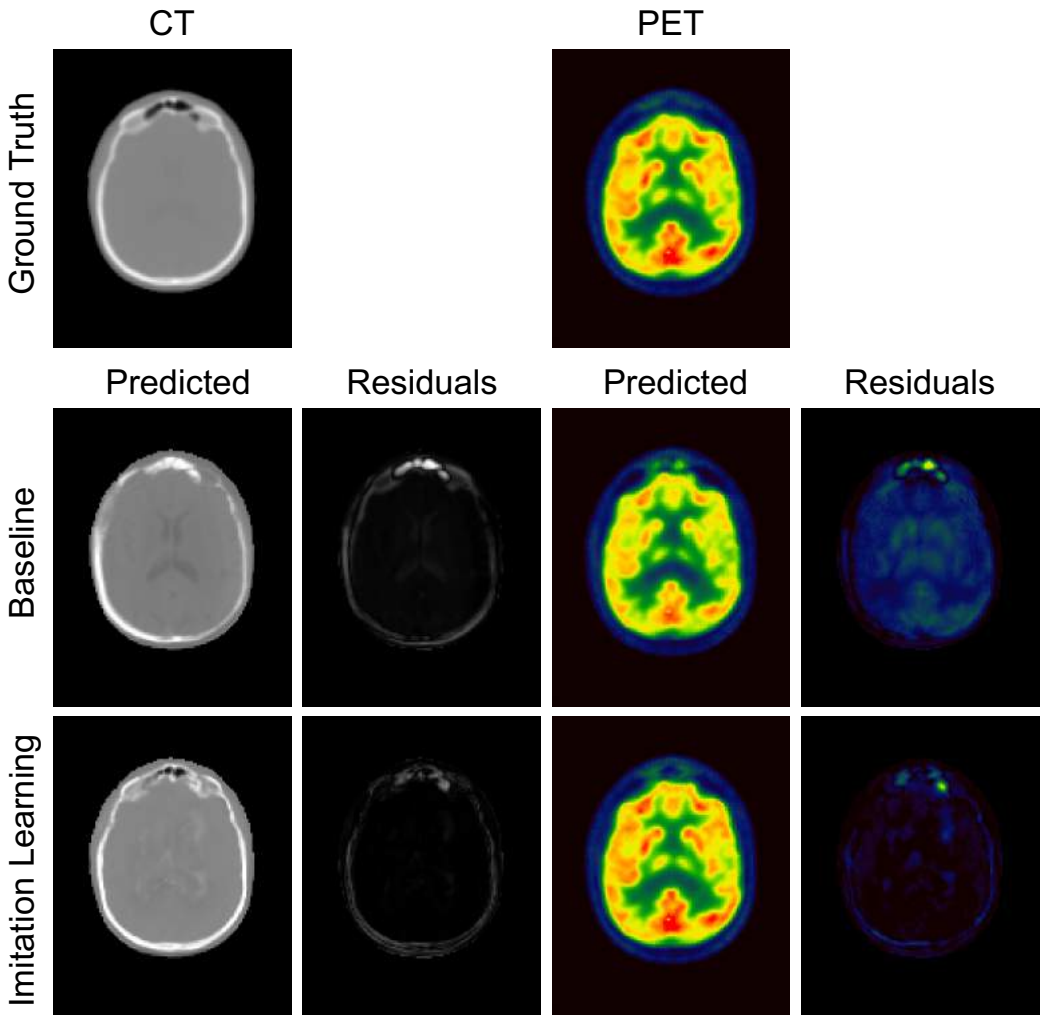

Figure 6.6: Qualitative results. From top to bottom: ground truth, baseline (HighRes3DNet) and imitation learning. From left to right: CT, pCT-CT residuals, PET, pPET-PET residuals. The error in the pCT generated with the proposed imitation learning is lower than the baseline pseudo CT residuals. The error in the pPET reconstructed with the proposed method is significantly lower than the pseudo PET error for the baseline method.

method, multi-atlas propagated pseudo CT and pseudo CT generated with proposed method and corresponding reconstructed PET images are presented in Fig. 6.7 for three subjects whose pseudo PET showed the lowest, the average, and the highest MAE.

Lastly, both the pseudo CT images and the pseudo PET images were mapped to a common space following a CT-based groupwise registration method (Rohlfing et al. 2001). The average across all subjects of the absolute pseudo CT error map was computed and the absolute pseudo PET error map (Fig. 6.8 top). Note that the average error in the pseudo CT for all three methods is centered in the skull region and only shows small improvement for the pseudo CT generated with the proposed imitation learning. However, looking at

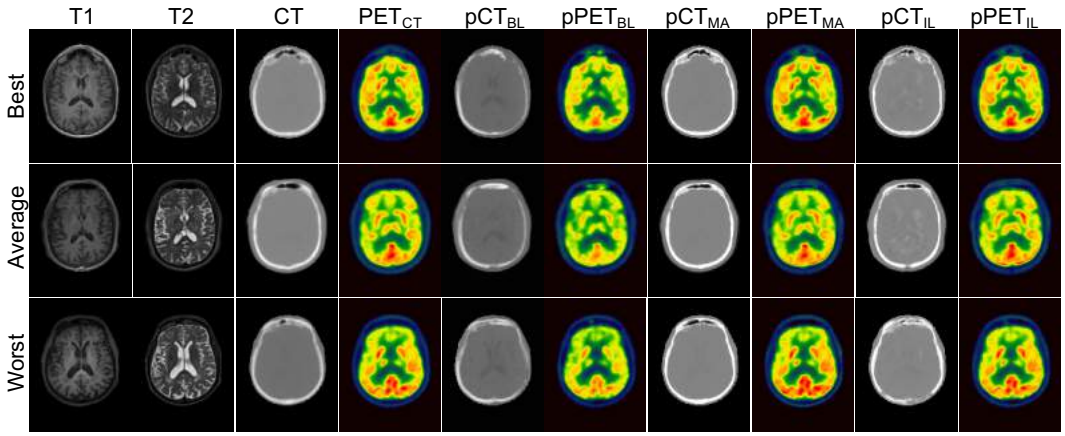

Figure 6.7: From left to right: the acquired T1-, T2-weighted MR, CT, and ground truth ^{18}F-FDG PET, the pseudo CT and pseudo PET generated with the baseline (HighRes3DNet only), the pseudo CT and pseudo PET generated with the multi-atlas propagation, and the pseudo CT, and pseudo PET generated with the proposed imitation learning for the subjects within the independent validation dataset that obtained the lowest (top row), average (middle row), and highest (bottom row) MAE in the pseudo PET.

the absolute difference of the pseudo PET and the gold standard PET, it can be seen that the average uptake error in the pseudo PET reconstructed with the baseline pseudo CT is significantly higher than in the pseudo PET reconstructed with the pseudo CT synthesized with the proposed imitation learning. Furthermore, it can be observed that small intensity differences in the skull region in the pseudo CT generated with the multi-atlas propagation method cause a significantly higher uptake error in the pseudo PET when this pseudo CT is used for pseudo PET reconstruction. The bottom row of Fig. 6.8 shows the standard deviation across all 23 subjects of pseudo CT and pseudo PET difference maps. It is noticeable that the standard deviation in the average pseudo CT error map is smaller for the proposed method compared to the baseline and the multi-atlas propagation method. Furthermore, the standard deviation of the groupwise average pseudo PET error is significantly higher for the pseudo PET difference map that was computed between the pseudo PET reconstructed with the baseline method and the gold standard PET compared to the pseudo PET difference map that was generated between the pseudo PET reconstructed with the proposed imitation learning method and the gold standard PET.

6.8 Discussion and conclusion

Following the hypothesis that the classical $\mathcal{L}_2$-loss is not necessarily the optimal minimization metric for CT synthesis, the presented multi-stage imitation learning network mini-

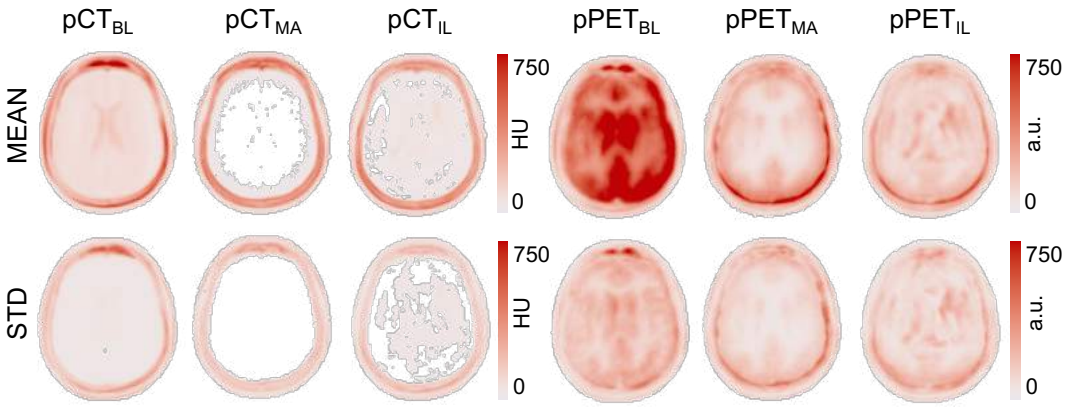

Figure 6.8: Groupwise average over 23 subjects (top) and standard deviation (bottom) of the pseudo CT absolute residuals of baseline, multi-atlas propagation and imitation learning (column 1-3) and pseudo PET absolute residuals between gold-standard PET and pseudo PETs reconstructed with baseline pseudo CT, multi-atlas propagation pseudo CT and imitation learning pseudoCT (column 4-6).

mizes a combination (as in Eq. 6.8) of the pixel-wise error between pseudo CT and CT and a proposed metric-loss that itself is represented by a CNN explicitly aiming at PET reconstruction application.

Two separate datasets were used in this chapter; one for training and cross-validation and another completely independent dataset to evaluate the performance of the proposed method on input images that were acquired with a different imaging protocol. The performance of the proposed imitation learning framework was compared to a feed forward network for pseudo CT synthesis that minimizes the classical $\mathcal{L}_2$-loss (HighRes3DNet). The results of the five-fold cross-validation in Table 6.1 demonstrate that the mean absolute error between the generated pseudo CT and the acquired ground truth CT is significantly lower compared to the baseline method for each fold. This observation leads to the hypothesis that this is likely caused by the regularizing nature of the imitation learning loss as all networks were trained until convergence. It can further be noted that the standard deviation for the proposed method is generally lower than the standard deviation of the baseline method. The lower error in the pseudo CT images subsequently results in a lower error in the reconstructed pseudo PET image when the pseudo CT is used as attenuation map for the PET reconstruction. The MAE in the whole head region and in the brain region only is significantly lower for the pseudo PET reconstructed with the proposed pseudo CT compared to the baseline pseudo CT. Difference images in Fig. 6.6 reveal that the errors in the pseudo CT are concentrated in the skull area, especially in areas with air/bone and soft-tissue/bone

boundaries like the nasal cavities. The wrongly predicted intensities in the skull region lead to incorrect attenuation maps that in turn lead to an overall underestimation of radionuclide uptake in the reconstructed pseudo PET images as shown in Fig. 6.6.

Quantitative results on a completely independent validation dataset are presented in Table 6.2 and confirm the improved performance of the proposed imitation learning network. The validation on the independent dataset was further extended by an additional comparison to a multi-atlas propagation method from Burgos et al. (Burgos et al. 2013) that is robust to image domain shifts. Results show that the error of the proposed pseudo CT lies around 111 HU whereas the baseline pseudo CT error is around 172 HU, which shows an improvement of approximately 35%. Even though a smaller pseudo CT error was not necessarily the aim of this work, the introduction of the imitation learning method has resulted in a better optimum and more generalizable model on average, which in this study resulted in an overall lower CT error compared to the baseline method. While the CT error was lower on average for the proposed method, there have been cases where only the PET error was lower compared to the baseline while the CT error was higher. This shows that inaccuracies in the CT are tolerable and do not compensate the performance of the proposed method when validated on the PET reconstruction accuracy. This is particularly obvious in Fig. 6.7, where synthesized pseudo CT images present multiple imperfections. However, those imperfections can be accepted, because the pseudo CT acts as interim step in the PET reconstruction process. Pseudo CT inaccuracies are thus a trade-off for better PET reconstruction. This method might not be well suited when trying to reconstruct a more realistic CT image.

Comparing the performance of the novel deep learning framework exploiting a combined pixel-wise and metric loss to the multi-atlas propagation method that is routinely used in clinical practice and clinical trial settings, the proposed method improves the pseudo CT synthesis performance by approximately 28%. The impact of the synthesis error in the pseudo CT on the pseudo PET is particularly present on the independent dataset that consisted of T1- and T2-weighted images that were acquired with a different imaging protocol than the training input MR data. The MAE in the pseudo PET reconstructed with the baseline is approximately 3.4 times higher and 1.5 times higher for the pseudo PET reconstructed with the multi-atlas propagated pseudo CT compared to the pseudo PET reconstructed with the proposed imitation learning pseudo CT (642.54 a.u. compared to

290.95 a.u. and 190.05 a.u., which represents 11%, 5% and 3% average uptake error respectively). Qualitative results in Fig. 6.7 illustrate the pseudo CTs and corresponding pseudo PETs of the independent validation and emphasize the underestimation of the skull in the baseline method and its missing ability to generate air/bone boundaries properly whereas pseudo CTs generated with the proposed method seem sharper than the ground truth CT images leading to pseudo PET images that reconstruct the radionuclide uptake more accurately. The pseudo CTs generated with the multi-atlas propagation method look visually sharper than the pseudo CTs generated with the imitation learning method, however, the density of the bone is overestimated which leads to an inaccurately radionuclide uptake in the reconstructed pseudo PET.

Analyzing the groupwise average difference and standard deviation across all 23 subjects of the independent dataset shows a similar performance on the pseudo CT synthesis for baseline, multi-atlas propagation and proposed imitation learning method as demonstrated in Fig. 6.8. However, when exploiting the average error map of the reconstructed radionuclide uptake the baseline method shows a significantly higher uptake error particularly in the brain region compared with the other two methods. The higher average difference in the skull region of the pseudo CT generated with the multi-atlas propagation method leads to a higher average error in the resulting pseudo PET image especially close to the skull. All three attenuation correction methods introduce a bias but the variance of the bias is lower when the pseudo PET is corrected with the attenuation map derived from the imitation learning pseudo CT.

The results of the validation on the independent dataset show a common problem of deep learning methods: image domain shift. Often, methods are developed to serve a problem specific purpose making them less generalizable, i.e., testing on images that are from a slightly different domain (in this case different MR acquisition protocols) than the training data fails. Multi-atlas propagation methods can overcome this problem since they rely on structural similarities in the image rather than voxel-wise intensity similarities. The proposed method shows to have good extrapolation properties due to a more realistic metric, which leads to less domain shift issues and an improved performance.

In this chapter it was shown that minimizing a combined loss that consists partly of the classical $\mathcal{L}_2$-loss and partly of a learned metric loss that itself minimizes the error in the reconstructed pseudo PET when the pseudo CT is used as attenuation map can indeed

significantly improve the PET reconstruction accuracy.

As a consequence of the newly introduced imitation learning loss, the performance of the pseudo CT synthesis on an image-based level was improved when optimizing the propsed network not only for the pseudo CT but also the pseudo PET error. Since the optimization happens over a high-dimensional model in a deep learning scenarios the introduction of the imitation loss appears to regularize the optimization function landscape better.

However, supervised deep learning based methods for pseudo CT synthesis for the purpose of PET/MR attenuation correction also have limitations relying on a co-registered database that represents a wide range of the population's anatomy. Small inaccuracies in the registration quality of the MR/CT database can have an influence on the training success. But, when validating on a database of images acquired with a different imaging protocol, the proposed end-to-end optimization strategy is robust enough to sustain local registration inaccuracies and acquisition protocol changes generating pseudo CT images that are significantly better than methods used in clinical practice. After all, accurately aligning CT and MR images is inevitable in order to validate the pixel-wise performance of any image synthesis algorithm until other appropriate methods have been developed that allow the validation of non-registered data.

Current limitations of the method arise from limited anatomical information in CT and MR images such as tumors as well as the tracer specificity of the proposed model. A larger database containing subjects with anatomical abnormalities could improve the robustness of the model. An uncertainty measure of the pseudo CT prediction (similar to chapter 5) could be integrated in the network providing a means of safety checking. This would make the method robust for clinical use by declining predictions that are highly uncertain if any extreme abnormalities in the input MR image are present that could cause the model to fail.

Compared to the results from the previously proposed DBR method, the end-to-end optimization presented in this chapter achieves inferior results. However, the end-to-end optimization strategy used HighRes3DNet as the convolutional neural network of choice when training both imitation and metric network, but can be exchanged with any other network architecture. Thus, future work could include to train the end-to-end optimization framework with DBR as network block that has shown to outperform a simple feed forward network such as HighRes3DNet. A combination of residual learning and imitation learning

has the potential to further increase the synthesis performance. While DBR has shown to reduce the CT residuals significantly, the end-to-end optimization further reduces the PET residuals. This way the strengths of both networks could be combined in future experiments. Additionally, recently trending deep learning techniques such as attention learning could be incorporated into the framework. Attention learning tries to imitate cognitive attention, which emphasizes the most important parts of the data. This way, attention could be used to emphasize the minimization of the PET reconstruction error. This idea could potentially even go further and include marginalization, where the contribution of each interim pseudo CT realisation towards the final attenuation corrected PET image would be considered in the model.

is a technique that mimics cognitive attention. The effect enhances the important parts of the input data and fades out the rest – the thought being that the network should devote more computing power on that small but important part of the data

In summary, this chapter presents a novel network architecture for pseudo CT synthesis in 3D for the purpose of PET/MR attenuation correction. Compared to state-of-the-art image synthesis CNNs, the proposed method does not assume the $\mathcal{L}_2$-loss, that is commonly used as a minimization metric in CT synthesis methods, as optimal when the ultimate aim is a low error in the corresponding pseudo PET when used as μ-map. Quantitative analysis on an out-of-distribution dataset shows that minimizing a more suitable metric that indeed optimizes for PET residuals (from CTs and pseudo CTs) can improve the process of CT synthesis for PET/MR attenuation correction. Furthermore, the proposed method proved to be robust to changes in the imaging protocol of the input T1- and T2-weighted MR images. Overall the proposed method provides a significant improvement in PET reconstruction accuracy when compared to a simple feed forward network and a multi-atlas propagation approach.

Chapter 7

General Conclusions

7.1 Summary

PET/MR imaging has a promising future ahead, combining functional imaging with an excellent soft tissue contrast, however, routine clinical use is still limited due to the imperfect attenuation correction that causes a high PET reconstruction bias. Thus, the overall aim of this thesis was to improve PET/MR attenuation correction by employing deep learning algorithms that have seen a rise in popularity in the field of computer science. The developed methods aim to outperform current non deep learning based state-of-the-art methods such as multi-atlas propagation and reach comparable results to CT-based attenuation correction.

In order to achieve this goal, the first novel PET/MR attenuation correction method developed in this work consists of a deep learning framework that synthesizes pseudo CT images from input MR images. This deep MR to CT synthesis framework is trained in a supervised manner, thus relying on a database of co-registered MR and CT image pairs. It utilizes a technique called boosting, known from classic machine learning, that exploits the idea that a collection of weak learners build a strong learner in their entirety. This way, each model aims to compensate the weaknesses of its predecessors. In the proposed framework this is achieved by concatenating multiple CNNs, representing a weak learner each, that when trained together build a strong learner that predicts a more accurate pseudo CT. The method can be seen as a form of residual learning, where the residuals of an initial prediction are minimized by additional learners further down the stream. It was shown that the proposed Deep Boosted Regression network can achieve significantly better results compared to traditional multi-atlas propagation methods and deep neural networks that do not benefit from the additional boosting mechanism. The Deep Boosted Regression framework was validated using ^{18}F-FDG PET images and a four-fold random bootstrapped validation

has shown that the PET reconstruction error is 50% lower than in multi-atlas propagation methods. Furthermore, it was demonstrated that when dealing with multi-modal data, a combined input improves the synthesis accuracy. However, when only one modality is available, T2-weighted MR images provide a better means for MR-based attenuation correction compared to their T1-weighted counterpart. An idea to further improve synthesis results is to use AC specific MR sequences that are used for segmentation-based AC methods, such as Dixon or UTE, as additional input channel. The network could potentially benefit from the additional bone information present in such sequences.

The second objective of this work was to develop a novel CT synthesis network for whole-body applications, which, until present, has remained a largely unchartered territory. As a means of safety, the aim was to incorporate uncertainty estimations to be aware of the network's prediction confidence. The proposed MultiRes network is a learning scheme for multi-resolution MR to CT synthesis of the full body. It is an end-to-end multi-scale convolutional neural network that takes input patches from full-body MR images at three resolution levels as inputs to synthesize high resolution, realistic CT patches. MultiRes$_{unc}$ is a version of this model that also incorporates explicit heteroscedastic uncertainty modelling by casting the task likelihood probabilistically, and epistemic uncertainty estimation via traditional Monte Carlo dropout. By learning feature maps of different resolutions the proposed method reduces the pseudo CT synthesis error significantly compared to other deep convolutional networks that only learn the image context from high-resolution images. Incorporating uncertainty into the model results in a slight decrease in performance however this is outweighed by the important additional information provided by the uncertainty. This contribution is particularly important for clinical whole-body imaging as current methods provide insufficient attenuation correction. While standard deep learning methods achieve already good results in brain imaging, this is not the case for whole-body imaging, mainly due to the large image size that comes with a limited field of view for training neural networks. However, the MultiRes$_{unc}$ approach requires more validation for clinical studies. The dataset that was used for training did not account for enough pathologies that could be present restricting the models generalizability. Furthermore, demographic characteristics of the study cohort exclude children. Therefore, it is of utmost importance to acquire more whole-body data in order to capture anatomical differences in the study population and make the model more generalizable.

The most common way to optimize deep MR to CT algorithms is to minimize the error between the synthesized pseudo CT and the corresponding ground truth CT image, equivalent to minimizing the $\mathcal{L}_2$-loss. However, minimizing the $\mathcal{L}_2$-loss between ground truth CT and CT prediction fails to recognize the main aim of CT synthesis, when used for PET/MR attenuation correction, which is to generate a synthetic CT that, when used as attenuation map for PET reconstruction, makes the reconstructed PET as close as possible to the gold standard PET reconstructed with the true CT. Thus, the third objective was to develop a novel deep learning framework for MR to CT synthesis that directly minimizes the PET residuals when the pseudo CT is used for PET reconstruction. To do so, a novel MR to CT synthesis method is introduced that is composed of two separate CNNs. A first CNN synthesizes multiple valid CT predictions using multi-hypothesis learning and a second CNN uses imitation learning in order to predict the residuals between ground truth PETs and PETs reconstructed with each valid pseudo CT. By minimizing this new metric loss, the network learns to synthesize pseudo CT images that will ultimately result in pseudo PETs with lower residuals. It has been shown that minimizing a more suitable metric that indeed optimizes for PET residuals can improve the process of CT synthesis for PET/MR attenuation correction. The proposed method further proved to be robust to changes in the imaging protocol of the input T1- and T2-weighted MR images. Overall the concept of imitation learning for MR to CT synthesis provides a significant improvement in PET reconstruction accuracy when compared to simpler CNNs and a multi-atlas propagation approach. The last contribution to this thesis is particularly promising for brain MR attenuation correction as it corrects for the ultimate PET error. While it was shown that the PET reconstruction accuracy can indeed be improved on brain images and shows sufficient generalizability, clinical integration into systems requires additional validation. It is especially important to account for pathologies like tumors and other abnormalities into the training dataset in order to use it routinely in clinical practice.

7.2 Limitations

The proposed methods showed promising results and efforts can be made to possibly integrate them into clinical systems. However, there are also challenges that the proposed methods face. These include a lack of generalizability, limited training data as well as memory constraints. In general, deep learning algorithms benefit from a large amount of training data in order to capture the wider context between two image domains. The datasets

used in this work were significantly smaller compared to datasets used in image translation networks used in computer vision where thousands of images are available. A large amount of data allows the network to generalize better and capture as many realistic scenarios, here pathologies, as possible. Therefore, results in this work are limited in their generalizability and would benefit from a larger amount of training data. Additionally, all methods presented are expected to perform better subject to better hardware, thus larger input training patches. Larger inpupt patches allow networks to learn the context between two imaging domains from a larger field of view, which is especially important for the large images acquired with whole-body imaging. Furthermore, all proposed networks were trained with the commonly used $\mathcal{L}_2$-loss that minimizes the sum of all squared differences between the ground truth and the predicted value. However, it is not robust to large outliers. The model will minimize this single outlier since errors of common examples appear small compared to that single outlier. $\mathcal{L}_1$-losses are more robust to outliers, however might not be able to capture small details in the image. A possible solution is to train a combined loss that minimizes both, the absolute error as well as the sum of all squared differences.

Overall, all presented methods need further validation to be used in clinical practice. Integrating a new attenuation correction method on a clinical scanner requires the method to be validated in-depth and cover all possible pathologies/scenarios that can occur as part of routine examinations. Additionally, there must be a back-up in case a model fails in a particular case in order to ensure robustness to the system. As discussed in chapter 5, uncertainty is a potential way to flag if a prediction can be trusted or if the network fails to synthesize a pseudo CT with sufficient quality for PET reconstruction. While all methods presented in this thesis showed promising results, they describe proof of concepts, wherefore the validation was limited to the data availabile and must be more extensive in order to integrate them into commercial PET/MR systems.

7.2.1 Generalizability (domain shift problem)

The results of the validation on an independent dataset in Chapter 6 show a common problem of deep learning methods: Image domain shift. Often methods are developed to serve a problem-specific purpose making them less generalizable, i.e. testing on images that are from a slightly different domain than the training data, in this case due to a different MR acquisition protocol, fails. Multi-atlas propagation methods are able to overcome this problem since they rely on structural similarities in the image rather than voxel-wise intensity

similarities. This domain shift problematic applies to the proposed methods in chapter 3 and chapter 5 as both methods rely on the $\mathcal{L}_2$-loss and the image properties within the given database. On the contrary, the proposed end-to-end optimization framework shows to have good extrapolation properties due to a more realistic metric, that itself is learned, which leads to less domain shift issues and an improved synthesis performance. In order to proof the concept behind the end-to-end optimization method, the network architecture used for this framework was HighRes3DNet, although it is theoretically possible to exchange the whole network block of the optimization framework with the DBR network architecture or the MultiRes$_{unc}$ architecture. However, this is not insignificant from a software engineering point of view due to the extended GPU memory requirements of those methods.

7.2.2 Imitation learning for whole-body images

The proposed MultiRes$_{unc}$ network shows promising results for whole-body MR to CT synthesis tasks which up to now have rarely been tackled. The method requires additional validation, especially on PET images when the synthesiszed pseudo CT is used as attenuation map. This is particularly difficult as the uptake distribution of PET tracers is much more localized in the whole-body. PET tracers accumulate in metabolic active regions, which include the brain and cancers both primary and metastatic. Thus, it can be expected that PET tracers will bind to specific regions within the whole body and are not confined to the brain. Furthermore, the brain is surrounded by a consistent bone cage, the skull, whereas photons that are ejected in an organ such as the lung might or might not travel through bone. It is therefore of high importance to have a good understanding of the patient's anatomy in order to assign the right attenuation coefficients. The proposed imitation learning strategy has shown promising results on the brain and could be used to directly optimize the model using both CT and PET residuals. However, a larger dataset is necessary to conduct experiments like this. Furthermore, models trained with the proposed imitation learning framework are tracer specific, which means that all PET data was acquired with ^{18}F-FDG, thus the model optimizes only for this tracer.

7.3 Future research direction

There are multiple directions to continue this research project that explore uncharted territory. One idea is to combine the advantage of the robustness of multi-atlas propagation algorithms with a deep learning framework, similar to creating a prior for the network that

then itself compensates for the weaknesses of the multi-atlas propagation alogorithm, such as its inability to synthesize abnormalities. To circumvent the generalizability issue that deep learning methods trained on a specific database face, domain adaptation techniques could be employed such that it is possible to train an imitation learning network on non tracer specific PET data. Another possible idea is to deploy the field of reinforcement learning for pseudo CT synthesis. The next three subchapters discuss possible implementations of these ideas.

7.3.1 Integration of multi-atlas-propagation as prior

One idea to improve the generalizability of deep learning methods for MR to CT synthesis is to make use of the advantages of other non deep learning related methods and integrate them in a deep learning framework. Multi-atlas propagation methods have proved to be a robust tool for MR to CT translation and have been the method of choice for many years. A possible framework could include a combination of such a method with a convolutional neural network. A multi-atlas propagation method can create an initial pseudo CT estimate that in itself acts as a prior for the neural network. The neural network then tries to compentate for the lack of extrapolation of the multi-atlas propagation method.

However, this combination would have the downside of being rather slow. The registrations in multi-atlas propagation methods are time costly, which would increase by adding an additional neural network into the pipeline. If it is possible to decrease the time that multi-atlas propagation methods require, this approach would be feasible. This could be achieved by fast deep learning-based image registration approaches. All registrations within the multi-atlas-propagation approach could be replaced by a registration network resulting in a prior for a deep learning image synthesis framework. This could have the benefit of providing the synthesis network a quickstart into the training process as much smaller registration errors must be compensated in the synthesis process, which has the additional benefit of reducing the time to converge the synthesis network.

7.3.2 Domain adaptation for imitation learning

A possible idea for the future to solve the domain shift problem and to increase the robustness of the proposed models is to utilize domain adaptation techniques. Such techniques have the ability to apply a model that was trained on one domain (also called *source* domain) to another domain (also called *target* domain). The target domain is usually related to the source domain but is different, for example when a model is trained on T1-weighted

brain images in order to detect tumors whereas the target at testing time is to detect brain bleeds. In this example target and source domain are related as both look at MR images of the skull, but different due to the different pathologies they present. The main aim is to make models more generalizable. Applying a domain adaptation strategy to the problem that the end-to-end optimization model is trained on PET images acquired with one tracer only, would mean that it is possible to train a model on PET data acquired with ^{18}F-FDG and test it on PET images acquired with a different tracer such as ^{11}C-RAC that binds to specific dopamine receptors. However, although domain adaptation itself is a large field of research gaining wide popularity at the moment, this problem is not trivial to solve. Another idea to allow the end-to-end optimization to be suitable for clinical use is to train multiple models with different PET tracers. This is a more practical solution that would allow a clinician to set the tracer as a parameter in the acquisition protocol. The scanner will perform the MR to CT translation on an imitation learning model that was trained with the specific tracer resulting in a tracer-specific reconstructed PET image.

7.3.3 Reinforcement learning

One possible idea for a new end-to-end optimization framework is to make use of reinforcement learning (RL), one of three machine learning paradigms next to unsupervised and supervised learning. In an RL scenario an agent tries to reach a specific goal by taking actions that lead to a reward depending on how well it performed with respect to a set policy. The overall aim of the algorithm is to predict the best subsequent step in order to maximize the associated reward. The way the agent chooses its actions relies both on learning from experience and exploration of new tactics. Thus, the performance of the agent improves with each iteration through this feedback loop.

7.3.4 Policy gradients

In policy-based reinforcement learning a parametrized function is learned which provides a distribution over possible actions, a, given a particular state, s. This policy is denoted as $\pi_\theta(a,s)$ where θ represents the network parameters to be learned. A typical learning scenario would collect experience as states, actions and their corresponding rewards. In this case the episodes of experience are of size 1 which is equivalent to a 1-step Markov Decision Process (MDP). Consider a policy objective function $J(\theta)$ for which optimal parameters, θ^*, need to be found, which maximize J, $\theta^* = \arg\max_\theta J(\theta)$. In order to maximize J, its derivative must be taken and stochastic gradient descent applied. The derivative is provided

by the Policy Gradient Theorem (Sutton et al. 2000),

Theorem 1 *For any differentiable policy $\pi_\theta(s,a)$ and a long-term value $Q^{\pi_\theta}(s,a)$ the policy gradient is given by,*

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a) Q^{\pi_\theta}(s,a)]$$

For the given 1-step MDP, the long-term value $Q^{\pi_\theta}(s,a)$ can be replaced by the final (and only) reward, R. This means the Policy Gradient Theorem can be rewritten as follows,

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a) R] = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \nabla_\theta \log \pi_\theta(s,a) R_s^a \tag{7.1}$$

In order to learn an optimal policy $\pi_{\theta^*}(s,a)$ tuples (s,a,R) are gathered and θ is updated using the REINFORCE algorithm (Williams 1992).

7.3.5 Potential reinforcement learning CT synthesis framework

The MR to CT synthesis problem for the purpose of PET reconstruction can be reformulated as a reinforcement learning problem. Figure 7.1 shows a potential reinforcement learning CT synthesis framework and can be summarized like the following:

1. The state, s, is the MR for a particular subject.
2. The actions, a, are the generated pseudo CTs.
3. The reward, R, is the $\mathcal{L}_2$ between the pseudo PET and the real PET.
4. The environment is NiftyPET which takes in an action and returns a reward.

The policy gradients formulation considers an RL agent which for a state s, draws an action a from a policy distribution $\pi_\theta(s,a)$. In the MR to CT synthesis case, this policy distribution is provided by a convolutional neural network. It has been shown that using dropout as part of a neural network and drawing various samples at inference time can be used as an unbiased estimate of the model variance (Gal & Ghahramani 2016). M samples are taken from this distribution to obtain a pixel-wise estimate of the variance, $\hat{\sigma}^2$, and a pixel-wise estimate of the mean $\hat{\mu}$. The probability of an action, a, given a state, s, is then given by a Gaussian distribution over each pixel. This representation is convenient as calculating the derivative of the log in order to compute $\nabla_\theta J(\theta)$ is a simpler task. In order to combine the reinforcement learning loss with direct supervision between the pseudo CT and the CT, $-J(\theta)$ is minimized. This term is denoted as $\mathcal{L}_{RL}$. At each iteration the

total loss is a sum of the $\mathcal{L}_2$ between the pseudo CT and the real CT and this RL loss, i.e $\mathcal{L} = \theta\mathcal{L}_2^{CT} + \theta\mathcal{L}_{RL}$.

The REINFORCE algorithm suffers from high variance and many approaches have attempted to address this such as Actor-Critic methods and Deterministic Policy Gradients. The potential reinforcement learning CT synthesis approach could address this variance by treating the RL loss as an auxiliary task with the main task being the CT synthesis. This approach is inspired by work from Du et al. published in 2018 (Du et al. 2018). The cosine similarity between gradient updates is measured and the gradient is obtained using the following update rule, $\nabla\mathcal{L}_2^{CT} + \nabla\mathcal{L}_2^{CT}\max(0, \cos(\nabla\mathcal{L}_{RL}\nabla\mathcal{L}_2^{CT}))$.

This kind of CT synthesis framework would be the first attempt to solve a medical image synthesis problem with reinforcement learning and could open up the way for self-learning synthesis algorithms in the medical imaging domain.

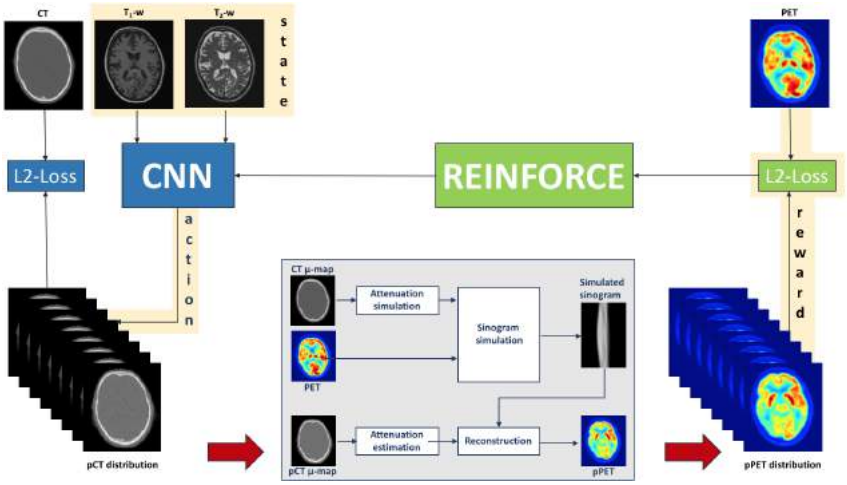

Figure 7.1: Potential reinforcement learning CT synthesis framework: T1- and T2-weighted MR images are fed into a CNN that synthesizes a Monte Carlo dropout distribution of pseudo CTs by minimizing the L_2-loss compared to the ground truth CT. A distribution of pseudo PETs is simulated with NiftyPET (grey box) by reconstructing simulated measured PET data using an attenuation map derived from each pseudo CT. The simulated measured PET data is generated by forward projecting the original PET using the Siemens mMR scanner geometry and multiplying the forward projected CT-based μ map. Pixel-wise reward distributions emerge from comparing pseudo PETs to original PETs that are an essential requirement for the REINFORCE algorithm that optimizes the policy gradient theorem in order to update the network's weights respectively.

Bibliography

Aitken, A. P., Giese, D., Tsoumpas, C., Schleyer, P., Kozerke, S., Prieto, C. & Schaeffter, T. (2014), 'Improved UTE-based attenuation correction for cranial PET-MR using dynamic magnetic field monitoring', *Medical Physics* **41(1)**.

Anazodo, U. C., Thiessen, J. D., Ssali, T., Mandel, J., Günther, M., Butler, J., Pavlosky, W., Prato, F. S., Thompson, R. T. & Lawrence, K. S. S. (2014), 'Feasibility of simultaneous whole-brain imaging on an integrated PET-MRI system using an enhanced 2-point Dixon attenuation correction method', *Frontiers in neuroscience* **8**.

Andreasen, D., Van Leemput, K., Hansen, R. H., Andersen, J. A. L. & Edmund, J. M. (2015), 'Patch-based generation of a pseudo CT from conventional MRI sequences for MRI-only radiotherapy of the brain', *Medical Physics* **42(4)**, 1596–1605.

Berker, Y., Franke, J., Salomon, A., Palmowski, M., Donker, H. C. W., Temur, Y., Mottaghy, F. M., Kuhl, C., Izquierdo-Garcia, D., Fayad, Z. A. et al. (2012), 'MRI-based attenuation correction for hybrid PET/MRI systems: a 4-class tissue segmentation technique using a combined ultrashort-echo-time/Dixon MRI sequence', *Journal of Nuclear Medicine* **53(5)**, 796–804.

Beyer, T., Townsend, D. W., Brun, T., Kinahan, P. E., Charron, M., Roddy, R., Jerin, J., Young, J., Byars, L., Nutt, R. et al. (2000), 'A combined pet/ct scanner for clinical oncology', *Journal of nuclear medicine* **41**(8), 1369–1379.

Brant-Zawadzki, M., Gillan, G. D. & Nitz, W. R. (1992), 'Mp rage: a three-dimensional, t1-weighted, gradient-echo sequence–initial experience in the brain.', *Radiology* **182**(3), 769–775.

Burger, C., Goerres, G., Schoenes, S., Buck, A., Lonn, A. & Von Schulthess, G. (2002), 'Pet attenuation coefficients from ct images: experimental evaluation of the transformation of

ct into pet 511-kev attenuation coefficients', *European journal of nuclear medicine and molecular imaging* **29**(7), 922–927.

Burgos, N., Cardoso, M. J., Modat, M., Pedemonte, S., Dickson, J., Barnes, A., Duncan, J. S., Atkinson, D., Arridge, S. R., Hutton, B. F. & Ourselin, S. (2013), Attenuation correction synthesis for hybrid PET-MR scanners, *in* 'Medical Image Computing and Computer-Assisted Intervention MICCAI 2013', pp. 147–154.

Burgos, N., Cardoso, M. J., Thielemans, K., Modat, M., Pedemonte, S., Dickson, J., Barnes, A., Ahmed, R., Mahoney, C. J., Schott, J. M., Duncan, J. S., Atkinson, D., Arridge, S. R., Hutton, B. F. & Ourselin, S. (2014), 'Attenuation Correction Synthesis for Hybrid PET-MR Scanners: Application to Brain Studies', *IEEE Transactions on Medical Imaging* **33(12)**, 2332–2341.

Cardoso, M. J., Sudre, C. H., Modat, M. & Ourselin, S. (2015), Template-based multimodal joint generative model of brain data, *in* 'International conference on information processing in medical imaging', Springer, pp. 17–29.

Catana, C., van der Kouwe, A., Benner, T., Michel, C. J., Hamm, M., Fenchel, M., Fischl, B., Rosen, B., Schmand, M. & Sorensen, A. G. (2010), 'Toward implementing an MRI-based PET attenuation-correction method for neurologic studies on the MR-PET brain prototype', *Journal of Nuclear Medicine* **51(9)**, 1431–1438.

Censor, Y., Gustafson, D. E., Lent, A. & Tuy, H. (1979), 'A new approach to the emission computerized tomography problem: Simultaneous calculation of attenuation and activity coefficients', *Nuclear Science, IEEE Transactions* **26(2)**, 2775–2779.

Chang, T., Diab, R. H., Clark Jr, J. W. & Mawlawi, O. R. (2013), 'Investigating the use of nonattenuation corrected PET images for the attenuation correction of PET data', *Medical Physics* **40(8)**.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. (2016), 3d u-net: learning dense volumetric segmentation from sparse annotation, *in* 'International conference on medical image computing and computer-assisted intervention', Springer, pp. 424–432.

Cohen, J. P., Luck, M. & Honari, S. (2018), Distribution matching losses can hallucinate features in medical image translation, *in* 'International conference on medical image computing and computer-assisted intervention', Springer, pp. 529–536.

Daftary, A. (2010), 'Pet-mri: Challenges and new directions', *Indian journal of nuclear medicine: IJNM: the official journal of the Society of Nuclear Medicine, India* **25**(1), 3.

Defrise, M., Rezaei, A. & Nuyts, J. (2012), 'Time-of-flight PET data determine the attenuation sinogram up to a constant', *Physics in Medicine and Biology* **57(4)**, 885.

Defrise, M., Rezaei, A. & Nuyts, J. (2014), 'Transmission-less attenuation correction in time-of-flight PET: Analysis of a discrete iterative algorithm', *Physics in Medicine and Biology* **59(4)**, 1073.

Dixon, W. T. (1984), 'Simple proton spectroscopic imaging', *Radiology* **153(1)**, 89–194.

Dong, X., Wang, T., Lei, Y., Higgins, K., Liu, T., Curran, W. J., Mao, H., Nye, J. A. & Yang, X. (2019), 'Synthetic ct generation from non-attenuation corrected pet images for whole-body pet imaging', *Physics in Medicine & Biology* **64**(21), 215016.

Dowling, J. A., Lambert, J., Parker, J., Salvado, O., Fripp, J., Capp, A., Wratten, C., Denham, J. W. & Greer, P. B. (2012), 'An atlas-based electron density mapping method for magnetic resonance imaging (MRI)-alone treatment planning and adaptive MRI-based prostate radiation therapy', *International Journal of Radiation Oncology* Biology* Physics* **83(1)**, e5–e11.

Du, Y., Czarnecki, W. M., Jayakumar, S. M., Pascanu, R. & Lakshminarayanan, B. (2018), 'Adapting auxiliary losses using gradient similarity', *arXiv preprint arXiv:1812.02224* .

Emami, H., Dong, M., Nejad-Davarani, S. P. & Glide-Hurst, C. K. (2018), 'Generating synthetic cts from magnetic resonance images using generative adversarial networks', *Medical physics* **45**(8), 3627–3636.

Gal, Y. & Ghahramani, Z. (2016), Dropout as a bayesian approximation: Representing model uncertainty in deep learning, *in* 'international conference on machine learning', pp. 1050–1059.

Ge, Y., Xue, Z., Cao, T. & Liao, S. (2019), Unpaired whole-body mr to ct synthesis with correlation coefficient constrained adversarial learning, *in* 'Medical Imaging 2019: Image Processing', Vol. 10949, International Society for Optics and Photonics, p. 1094905.

Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D. I., Wang, G., Eaton-Rosen, Z., Gray, R., Doel, T., Hu, Y. et al. (2018), 'Niftynet: a deep-learning platform for medical imaging', *Computer methods and programs in biomedicine* **158**, 113–122.

Glorot, X., Bordes, A. & Bengio, Y. (2011), Deep sparse rectifier neural networks, *in* 'Proceedings of the fourteenth international conference on artificial intelligence and statistics', pp. 315–323.

Gudur, M. S. R., Hara, W., Le, Q.-T., Wang, L., Xing, L. & Li, R. (2014), 'A unifying probabilistic bayesian approach to derive electron density from mri for radiation therapy treatment planning', *Physics in Medicine & Biology* **59**(21), 6595.

Han, X. (2017), 'Mr-based synthetic ct generation using a deep convolutional neural network method', *Medical physics* **44**(4), 1408–1419.

He, K., Zhang, X., Ren, S. & Sun, J. (2015*a*), 'Deep residual learning for image recognition', *CoRR* **abs/1512.03385**.
URL: *http://arxiv.org/abs/1512.03385*

He, K., Zhang, X., Ren, S. & Sun, J. (2015*b*), Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, *in* 'Proceedings of the IEEE international conference on computer vision', pp. 1026–1034.

Hiasa, Y., Otake, Y., Takao, M., Matsuoka, T., Takashima, K., Carass, A., Prince, J. L., Sugano, N. & Sato, Y. (2018), Cross-modality image synthesis from unpaired data using cyclegan, *in* 'International workshop on simulation and synthesis in medical imaging', Springer, pp. 31–41.

Hou, S. & Wang, Z. (2019), Weighted channel dropout for regularization of deep convolutional neural network, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 33, pp. 8425–8432.

Hsu, S.-H., Cao, Y., Huang, K., Feng, M. & Balter, J. M. (2013), 'Investigation of a method

for generating synthetic CT models from MRI scans of the head and neck for radiation therapy', *Physics in Medicine and Biology* **58(23)**, 8419.

Huo, Y., Xu, Z., Bao, S., Assad, A., Abramson, R. G. & Landman, B. A. (2018), Adversarial synthesis learning enables segmentation without target modality ground truth, *in* '2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)', IEEE, pp. 1217–1220.

Huynh, T., Gao, Y., Kang, J., Wang, L., Zhang, P., Lian, J. & Shen, D. (2016), 'Estimating CT image from MRI data using structured random forest and auto-context model', *IEEE Transactions on Medical Imaging* **35(1)**, 174–183.

Hwang, D., Kang, S. K., Kim, K. Y., Seo, S., Paeng, J. C., Lee, D. S. & Lee, J. S. (2019), 'Generation of pet attenuation map for whole-body time-of-flight 18f-fdg pet/mri using a deep neural network trained with simultaneously reconstructed activity and attenuation maps', *Journal of Nuclear Medicine* **60**(8), 1183–1189.

Hwang, D., Kim, K. Y., Kang, S. K., Seo, S., Paeng, J. C., Lee, D. S. & Lee, J. S. (2018), 'Improving the accuracy of simultaneously reconstructed activity and attenuation maps using deep learning', *Journal of Nuclear Medicine* **59**(10), 1624–1629.

Izquierdo-Garcia, D., Hansen, A. E., Förster, S., Benoit, D., Schachoff, S., Fürst, S., Chen, K. T., Chonde, D. B. & Catana, C. (2014), 'An SPM8-based approach for attenuation correction combining segmentation and nonrigid template formation: application to simultaneous PET/MR brain imaging', *Journal of Nuclear Medicine* **55(11)**, 1825–1830.

Johansson, A., Karlsson, M. & Nyholm, T. (2011), 'CT substitute derived from MRI sequences with ultrashort echo time', *Medical physics* **38(5)**, 2708–2714.

Jonsson, J. H., Akhtari, M. M., Karlsson, M. G., Johansson, A., Asklund, T. & Nyholm, T. (2015), 'Accuracy of inverse treatment planning on substitute ct images derived from mr data for brain lesions', *Radiation Oncology* **10**(1), 1–7.

Juttukonda, M. R., Mersereau, B. G., Chen, Y., Su, Y., Rubin, B. G., Benzinger, T. L. S., Lalush, D. S. & An, H. (2015), 'MR-based attenuation correction for PET/MRI neurological studies with continuous-valued attenuation coefficients for bone through a conversion from R2* to CT-Hounsfield units', *Neuroimage* **112**, 160–168.

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D. & Glocker, B. (2017), 'Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation', *Medical image analysis* **36**, 61–78.

Kawaguchi, H., Hirano, Y., Y., E., Kershaw, J., Shiraishi, T., Suga, M., Ikoma, Y., Obata, T.and Ito, H. & Yamaya, T. (2014), 'A proposal for PET/MRI attenuation correction with m-values measured using a fixed-position radiation source and MRI segmentation', *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **734**, 156–161.

Keereman, V., Fierens, Y., Broux, T., De Deene, Y., Lonneux, M. & Vandenberghe, S. (2010), 'MRI-based attenuation correction for PET/MRI using ultrashort echo time sequences', *Journal of Nuclear Medicine* **51(5)**, 812–818.

Kingma, D. P. & Ba, J. (2014), 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980* .

Kläser, K., Borges, P., Shaw, R., Ranzini, M., Modat, M., Atkinson, D., Thielemans, K., Hutton, B. F., Goh, V., Cook, G., Cardoso, M. J. & Ourselin, S. (2020), Uncertainty-aware multi-resolution whole-body mr to ct synthesis, *in* 'International Workshop on Simulation and Synthesis in Medical Imaging', Springer.

Kläser, K., Markiewicz, P., Ranzini, M., Li, W., Modat, M., Hutton, B. F., Atkinson, D., Thielemans, K., Cardoso, M. J. & Ourselin, S. (2018), Deep boosted regression for mr to ct synthesis, *in* 'International Workshop on Simulation and Synthesis in Medical Imaging', Springer, pp. 61–70.

Kops, E. R. & Herzog, H. (2007), Alternative methods for attenuation correction for PET images in MR-PET scanners, *in* 'Nuclear Science Symposium Conference Record, 2007. NSS'07. IEEE', Vol. 6, pp. 4327–4330.

Korhonen, J., Kapanen, M., Keyriläinen, J., Seppälä, T. & Tenhunen, M. (2014), 'A dual model hu conversion from mri intensity values within and outside of bone segment for mri-based radiotherapy treatment planning of prostate cancer', *Medical physics* **41**(1), 011704.

Ladefoged, C. N., Benoit, D., Law, I., Holm, S., Kjær, A., Højgaard, L., Hansen, A. E. & Andersen, F. L. (2015), 'Region specific optimization of continuous linear attenuation coefficients based on UTE (RESOLUTE): application to PET/MR brain imaging', *Physics in Medicine and Biology* **60(20)**, 8047.

Ladefoged, C. N., Law, I., Anazodo, U., Lawrence, K. S., Izquierdo-Garcia, D., Catana, C., Burgos, N., Cardoso, M. J., Ourselin, S., Hutton, B. et al. (2017), 'A multi-centre evaluation of eleven clinically feasible brain pet/mri attenuation correction techniques using a large cohort of patients', *Neuroimage* **147**, 346–359.

Le Goff-Rougetet, R., Frouin, V., Mangin, J. F. & Bendriem, B. (1994), 'Segmented MR images for brain attenuation correction in PET', *Medical Imaging* pp. 725–736.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. & Jackel, L. D. (1989), 'Backpropagation applied to handwritten zip code recognition', *Neural computation* **1**(4), 541–551.

Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M. J. & Vercauteren, T. (2017), On the compactness, efficiency, and representation of 3d convolutional networks: brain parcellation as a pretext task, *in* 'International conference on information processing in medical imaging', Springer, pp. 348–360.

Markiewicz, P. J., Ehrhardt, M. J., Erlandsson, K., Noonan, P. J., Barnes, A., Schott, J. M., Atkinson, D., Arridge, S. R., Hutton, B. F. & Ourselin, S. (2018), 'Niftypet: a high-throughput software platform for high quantitative accuracy and precision pet imaging and analysis', *Neuroinformatics* **16**(1), 95–115.
URL: *https://doi.org/10.1007/s12021-017-9352-y*

Martinez-Möller, A., Souvatzoglou, M., Delso, G., Bundschuh, R. A., Chefd'hotel, C., Ziegler, S. I., Navab, N., Schwaiger, M. & Nekolla, S. G. (2009), 'Tissue classification as a potential approach for attenuation correction in whole-body PET/MRI: Evaluation with PET/CT data', *Journal of Nuclear Medicine* **50(4)**, 520–6.

Maspero, M., Savenije, M. H., Dinkla, A. M., Seevinck, P. R., Intven, M. P., Jurgenliemk-Schulz, I. M., Kerkmeijer, L. G. & van den Berg, C. A. (2018), 'Dose evaluation of fast synthetic-ct generation using a generative adversarial network for general pelvis mr-only radiotherapy', *Physics in Medicine & Biology* **63**(18), 185001.

Mehranian, A. & Zaidi, H. (2015), 'Emission-based estimation of lung attenuation coefficients for attenuation correction in time-of-flight PET/MR', *Physics in Medicine and Biology* **60(12)**, 4813.

Meikle, S. R., Bailey, D. L., Hooper, P. K., Eberl, S., Hutton, B. F., Jones, W. F., Fulton, R. R. & Fulham, M. J. (1995), 'Simultaneous emission and transmission measurements for attenuation correction in whole-body PET', *Journal of Nuclear Medicine* **36(9)**, 1680–1688.

Mérida, I., Costes, N., Heckemann, R. A., Drzezga, A., Förster, S. & Hammers, A. (2015), Evaluation of several multi-atlas methods for PSEUDO-CT generation in brain MRI-PET attenuation correction, *in* 'Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on', pp. 1431–1434.

Modat, M., Cash, D. M., Daga, P., Winston, G. P., Duncan, J. S. & Ourselin, S. (2014), A symmetric block-matching framework for global registration, *in* 'Medical Imaging 2014: Image Processing', Vol. 9034, International Society for Optics and Photonics, p. 90341D.

Modat, M., Ridgway, G. R., Taylor, Z. A., Lehmann, M., Barnes, J., Hawkes, D. J., Fox, N. C. & Ourselin, S. (2010), 'Fast free-form deformation using graphics processing units', *Computer methods and programs in biomedicine* **98**(3), 278–284.

Mollet, P., Keereman, V., Bini, J., Izquierdo-Garcia, D., Fayad, Z. A. & Vandenberghe, S. (2014), 'Improvement of attenuation correction in time-of-flight PET/MR imaging with a positron-emitting source', *Journal of Nuclear Medicine* **55(2)**, 329–336.

Mollet, P., Keereman, V., Clementel, E. & Vandenberghe, S. (2012), 'Simultaneous MR-compatible emission and transmission imaging for PET using time-of-flight information', *Medical Imaging, IEEE Transactions* **31(9)**, 1734–1742.

Natterer, F. & Herzog, H. (1992), 'Attenuation correction in positron emission tomography', *Mathematical Methods in the Applied Sciences* **15(5)**, 321–330.

Nie, D., Trullo, R., Lian, J., Wang, L., Petitjean, C., Ruan, S., Wang, Q. & Shen, D. (2018), 'Medical image synthesis with deep convolutional adversarial networks', *IEEE Transactions on Biomedical Engineering* **65**(12), 2720–2730.

Nie, D.and Cao, X., Gao, Y., Wang, L. & Shen, D. (2016), Estimating CT image from MRI data using 3D fully convolutional networks, *in* 'International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis', pp. 170–178.

Nuyts, J., Dupont, P., Stroobants, S., Benninck, R., Mortelmans, L. & Suetens, P. (1999), 'Simultaneous maximum a posteriori reconstruction of attenuation and activity distributions from emission sinograms', *Medical Imaging, IEEE Transactions* **18(5)**, 393–403.

Ourselin, S., Roche, A., Subsol, G., Pennec, X. & Ayache, N. (2001), 'Reconstructing a 3d structure from serial histological sections', *Image and vision computing* **19**(1-2), 25–31.

Rezaei, A., Defrise, M., Bal, G., Michel, C., Conti, M., Watson, C. & Nuyts, J. (2012*a*), 'Simultaneous reconstruction of activity and attenuation in time-of-flight PET', *Medical Imaging, IEEE Transactions* **31(12)**, 2224–2233.

Rezaei, A., Defrise, M., Bal, G., Michel, C., Conti, M., Watson, C. & Nuyts, J. (2012*b*), 'Simultaneous reconstruction of activity and attenuation in time-of-flight pet', *IEEE transactions on Medical Imaging* **31**(12), 2224–2233.

Rezaei, A., Defrise, M. & Nuyts, J. (2014), 'ML-reconstruction for TOF-PET with simultaneous estimation of the attenuation factors', *Medical Imaging, IEEE Transactions* **33(7)**, 1563–1572.

Rohlfing, T., Brandt, R., Maurer, C. R. & Menzel, R. (2001), Bee brains, b-splines and computational democracy: Generating an average shape atlas, *in* 'Proceedings IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA 2001)', IEEE, pp. 187–194.

Ronneberger, O., Fischer, P. & Brox, T. (2015), U-net: Convolutional networks for biomedical image segmentation, *in* 'International Conference on Medical image computing and computer-assisted intervention', Springer, pp. 234–241.

Roy, S., Wang, W.-T., Carass, A., Prince, J. L., Butman, J. A. & Pham, D. L. (2014), 'PET attenuation correction using synthetic CT from ultrashort echo-time MR imaging', *Journal of Nuclear Medicine* **55(12)**, 2071–2077.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986), 'Learning representations by back-propagating errors', *nature* **323**(6088), 533–536.

Rupprecht, C., Laina, I., DiPietro, R., Baust, M., Tombari, F., Navab, N. & Hager, G. D. (2017), Learning in an uncertain world: Representing ambiguity through multiple hypotheses, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 3591–3600.

Salomon, A., Goedicke, A., Schweizer, B., Aach, T. & Schulz, V. (2010), 'Simultaneous reconstruction of activity and attenuation for pet/mr', *IEEE transactions on medical imaging* **30**(3), 804–813.

Salomon, A., Goedicke, A., Schweizer, B., Aach, T. & Schulz, V. (2011), 'Simultaneous reconstruction of activity and attenuation for PET/MR', *Medical Imaging, IEEE Transactions* **30(3)**, 804–13.

Schapire, R. E. (1990), 'The strength of weak learnability', *Machine learning* **5**(2), 197–227.

Schleyer, P. J., Schaeffter, T. & Marsden, P. K. (2010), 'The effect of inaccurate bone attenuation coefficient and segmentation on reconstructed PET images', *Nuclear Medicine Communications* **31(8)**, 708–16.

Schreibmann, E., Nye, J. A., Schuster, D. M., Martin, D. R., Votaw, J. & Fox, T. (2010), 'MR-based attenuation correction for hybrid PET-MR brain imaging systems using deformable image registration', *Medical Physics* **37(5)**, 2101–2109.

Schulz, V., Torres-Espallardo, I., Renisch, S., Hu, Z., Ojha, N., Börnert, P., Perkuhn, M., Niendorf, T., Schäfer, W. M., Brockmann, H., Krohn, T., Buhl, A., Günther, R., Mottaghy, F. M. & Krombach, G. A. (2011), 'Automatic, three-segment, MR-based attenuation correction for whole-body PET/MR data', *European Journal of Nuclear Medicine and Molecular Imaging* **38(1)**, 138–152.

Sikka, A., Peri, S. V. & Bathula, D. R. (2018), Mri to fdg-pet: cross-modal synthesis using 3d u-net for multi-modal alzheimers classification, *in* 'International Workshop on Simulation and Synthesis in Medical Imaging', Springer, pp. 80–89.

Sjölund, J., Forsberg, D., Andersson, M. & Knutsson, H. (2015), 'Generating patient specific pseudo-CT of the head from MR using atlas-based regression', *Physics in Medicine and Biology* **60(2)**, 825.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. (2014), 'Dropout: a simple way to prevent neural networks from overfitting', *The journal of machine learning research* **15**(1), 1929–1958.

Su, K.-H., Hu, L., Stehning, C., Helle, M., Qian, P., Thompson, C. L., Pereira, G. C., Jordan, D. W., Herrmann, K. A., Traughber, M. et al. (2015), 'Generation of brain pseudo-CTs using an undersampled, single-acquisition UTE-mDixon pulse sequence and unsupervised clustering', *Medical Physics* **42(8)**, 4974–4986.

Sutton, R. S., McAllester, D. A., Singh, S. P. & Mansour, Y. (2000), Policy gradient methods for reinforcement learning with function approximation, *in* 'Advances in neural information processing systems', pp. 1057–1063.

Tissue substitutes in radiation dosimetry and measurement (1989).
URL: *https://www.osti.gov/biblio/10102048*

Townsend, D. W. (2008), 'Dual-modality imaging: combining anatomy and function', *Journal of Nuclear Medicine* **49**(6), 938–955.

Uh, J., Merchant, T. E., Li, Y., Li, X. & Hua, C. (2014), 'Mri-based treatment planning with pseudo ct generated through atlas registration', *Medical physics* **41**(5), 051711.

Wagenknecht, G., Kops, E. R., Tellmann, L. & Herzog, H. (2009), Knowledge-based segmentation of attenuation-relevant regions of the head in T1-weighted MR images for attenuation correction in MR/PET systems, *in* 'Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE', IEEE, pp. 3338–3343.

Williams, R. J. (1992), 'Simple statistical gradient-following algorithms for connectionist reinforcement learning', *Machine learning* **8**(3-4), 229–256.

Wolterink, J. M., Dinkla, A. M., Savenije, M. H., Seevinck, P. R., van den Berg, C. A. & Išgum, I. (2017), Deep mr to ct synthesis using unpaired data, *in* 'International workshop on simulation and synthesis in medical imaging', Springer, pp. 14–23.

Yaakub, S. N., McGinnity, C. J., Clough, J. R., Kerfoot, E., Girard, N., Guedj, E. & Hammers, A. (2019), Pseudo-normal pet synthesis with generative adversarial networks for localising hypometabolism in epilepsies, *in* 'International Workshop on Simulation and Synthesis in Medical Imaging', Springer, pp. 42–51.

Yang, H. et al. (2018), Unpaired brain mr-to-ct synthesis using a structure-constrained cyclegan, *in* 'Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support', Springer International Publishing, pp. 174–182.

Yang, X. & Fei, B. (2013), 'Multiscale segmentation of the skull in MR images for MRI-based attenuation correction of combined MR/PET', *Journal of the American Medical Informatics Association* **20(6)**, 1037–1045.

Zaidi, H., Montandon, M.-L. & Slosman, D. O. (2003), 'Magnetic resonance imaging-guided attenuation and scatter corrections in three-dimensional brain positron emission tomography', *Medical Physics* **30(5)**, 937–948.

Zhang, Z., Yang, L. & Zheng, Y. (2018), Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network, *in* 'Proceedings of the IEEE conference on computer vision and pattern Recognition', pp. 9242–9251.

Zhao, Y., Liao, S., Guo, Y., Zhao, L., Yan, Z., Hong, S., Hermosillo, G., Liu, T., Zhou, X. S. & Zhan, Y. (2018), Towards mr-only radiotherapy treatment planning: synthetic ct generation using multi-view deep convolutional neural networks, *in* 'International Conference on Medical Image Computing and Computer-Assisted Intervention', Springer, pp. 286–294.

Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. (2017), Unpaired image-to-image translation using cycle-consistent adversarial networks, *in* 'Proceedings of the IEEE international conference on computer vision', pp. 2223–2232.