Prior Knowledge for Deep Learning Based Interventional Cone Beam Computed Tomography Reconstruction

DISSERTATION

zur Erlangung des akademischen Grades

Doktoringenieur (Dr.-Ing.)

angenommen durch die Fakultät für Informatik
der Otto-von-Guericke-Universität Magdeburg

von MSc Philipp Ernst

geb. am 22.06.1995 in Ebersbach/Sa.

Gutachterinnen/Gutachter

Prof. Dr.-Ing. Andreas Nürnberger
Prof. Dr. rer. nat. Georg Rose
Prof. Dr. Giuseppe Placidi
Eingereicht am 23.08.2022

Magdeburg, den 05.05.2023

Ehrenerklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; verwendete fremde und eigene Quellen sind als solche kenntlich gemacht. Insbesondere habe ich nicht die Hilfe eines kommerziellen Promotionsberaters in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Ich habe insbesondere nicht wissentlich:

- Ergebnisse erfunden oder widersprüchliche Ergebnisse verschwiegen,
- statistische Verfahren absichtlich missbraucht, um Daten in ungerechtfertigter Weise zu interpretieren,
- fremde Ergebnisse oder Veröffentlichungen plagiiert,
- fremde Forschungsergebnisse verzerrt wiedergegeben.

Mir ist bekannt, dass Verstöße gegen das Urheberrecht Unterlassungs- und Schadensersatzansprüche des Urhebers sowie eine strafrechtliche Ahndung durch die Strafverfolgungsbehörden begründen kann. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form als Dissertation eingereicht und ist als Ganzes auch noch nicht veröffentlicht.

Magdeburg, den 23.08.2022

Philipp Ernst

Abstract

Computed tomography (CT) is a non-invasive imaging method for anatomical structures. Due to its fast acquisition time, it is also suitable for surgery in near real-time using interventional C-arm devices. However, the biggest downside of CT is the harmful X-radiation that this imaging method is based on and that the patient as well as the surgeons are exposed to. The dose can be reduced by acquiring fewer X-ray projections or by reducing the current of the X-ray tube, which leads to streaking artifacts or noisy reconstructions, respectively.

Deep learning has become one of the major machine learning methods in the past decade and also found its way into medical imaging. Compared to traditional approaches, deep learning and convolutional neural networks (CNNs) are usually very fast and superior in accuracy. For this reason, the aim of this thesis is to investigate how CNNs can be applied to interventional CT to reduce the amount of radiation while maintaining the image quality of the reconstructions.

Several studies have shown that, despite training plain CNNs already improves the image quality, providing more specific, task-related information is beneficial for guiding neural networks to solutions that resemble the sought outcomes closer. In this thesis, the influence of this prior knowledge for the task of optimizing sparse view CT reconstructions is investigated systematically by categorizing it into the three classes *Algebraic*, *Machine Learning* and *Temporal and Model Prior Knowledge*. Examples of each category are explained and investigated, it is discussed if and how the categories can be combined to accumulate and concentrate the information, how the models perform with respect to different numbers of acquired projections, and which evaluation metrics are suited best for assessing the quality of the optimized CT images.

Moreover, failed attempts are presented and possible reasons and explanations for why the initial hypotheses had to be rejected are explored.

These studies are going to give a deeper insight into what information is essential and beneficial for interventional CT reconstructions using deep learning architectures. Further analyses will show to what extent the developed models can be applied in clinical practice and which limitations have to be dealt with.

Zusammenfassung

Die Computertomographie (CT) ist ein nicht-invasives Bildgebungsverfahren für anatomische Strukturen. Aufgrund ihrer schnellen Aufnahmezeit eignet sie sich auch für chirurgische Eingriffe in nahezu Echtzeit mit interventionellen C-Bogen-Geräten. Der größte Nachteil der CT ist jedoch die schädliche Röntgenstrahlung, auf der diese Bildgebungsmethode beruht und der sowohl der Patient als auch die Chirurgen ausgesetzt sind. Die Dosis kann reduziert werden, indem weniger Röntgenprojektionen aufgenommen werden oder die Stromstärke der Röntgenröhre reduziert wird, was zu Streifenartefakten bzw. verrauschten Rekonstruktionen führt.

Deep Learning hat sich im letzten Jahrzehnt zu einer der wichtigsten Methoden des maschinellen Lernens entwickelt und auch in der medizinischen Bildgebung Einzug gehalten. Im Vergleich zu traditionellen Ansätzen sind Deep Learning und Convolutional Neural Networks (CNNs) in der Regel sehr schnell und in ihrer Genauigkeit überlegen. Aus diesem Grund soll in dieser Arbeit untersucht werden, wie CNNs in der interventionellen CT eingesetzt werden können, um die Strahlenbelastung zu reduzieren und gleichzeitig die Bildqualität der Rekonstruktionen zu erhalten.

Studien haben gezeigt, dass, obwohl das Training einfacher CNNs bereits die Bildqualität verbessert, die Bereitstellung spezifischerer, aufgabenbezogener Informationen von Vorteil ist, um neuronale Netze zu Lösungen zu führen, die den angestrebten Ergebnissen näher kommen. In dieser Arbeit wird der Einfluss dieses Vorwissens für die Aufgabe der Optimierung von CT-Rekonstruktionen aus wenigen Projektionen systematisch untersucht, indem es in die drei Klassen *Algebraisches*, *Machine Learning* und *Temporales und Modellvorwissen* eingeteilt wird. Beispiele für jede Kategorie werden erläutert und untersucht, es wird diskutiert, ob und wie die Kategorien kombiniert werden können, um die Informationen zu akkumulieren und zu konzentrieren, wie die Qualität der Modelle mit der Anzahl der erfassten Projektionen korreliert und welche Bewertungsmetriken sich am besten für die Beurteilung der Qualität der optimierten CT-Bilder eignen. Darüber hinaus werden gescheiterte Versuche vorgestellt und mögliche Gründe und Erklärungen dafür erforscht, warum die ursprünglichen Hypothesen verworfen werden mussten.

Diese Studien geben einen tieferen Einblick, welche Informationen für interventionelle CT-Rekonstruktionen unter Verwendung von Deep-Learning-Architekturen wesentlich und nützlich sind. Weitere Analysen zeigen, inwieweit die entwickelten Modelle in der klinischen Praxis angewendet werden können und welche Einschränkungen zu beachten sind.

Contents

Acronyms

CT	Computed Tomography
iCT	Interventional CT
CBCT	Cone Beam CT
USCT	Ultrasound CT
MRI	Magnetic Resonance Imaging
FBP	Filtered Backprojection
FDK	Feldkamp-Davis-Kress
ART	Algebraic Reconstruction Technique
CNN	Convolutional Neural Network
DBP	Differentiated Backprojection
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
NMSE	Normalized Mean Squared Error
NRMSE	Normalizaed Root Mean Squared Error
PSNR	Peak Signal-to-Noise Ratio
MAE	Mean Absolute Error
SSIM	Structural Similarity Image Measure

Chapter 1

Introduction

This chapter motivates the general topic of the thesis, provides a definition for (the types of) prior knowledge, and points out the research questions. Finally, the significance and the limitations of the research are described.

1.1 Motivation

Cancer is one of the most prevalent diseases of the human body. It is characterized by abnormal cell growth leading to malignant tumors which destroy surrounding tissue and may form metastases, i.e. spread to different sites inside the body. According to the Council of the European Union, one out of three people in Europe develop some kind of cancer in their life [CK08]. In 2020, almost twenty million new cases were diagnosed globally, which corresponds to more than 50,000 diagnoses per day. Not only is cancer very prevalent but also associated with a high mortality: almost ten million deaths due to cancer were registered globally in 2020 [SFS+21].

The incidences for the different types of cancer are not equal for males and females. Lung and prostate cancer are the most common kinds of cancer for men (14.3 % and 14.1 %), while breast cancer is the most commonly diagnosed type for women (24.5 %).

Luckily, there are many possible treatments specific to the different cancer types, especially if diagnosed in an early stage of tumor growth and when no metastases have formed yet. To this end, cancer screening is an essential part of the diagnosis. For the previously mentioned most common cancer types, this includes mammography for breast cancer, low dose Computed Tomography (CT) or chest radiographs for lung cancer, and digital rectal examination or performing prostate-specific antigen measurements in the blood for prostate cancer.

Once there is an initial suspicion, the definite diagnosis is usually made with a biopsy and subsequent clinical analysis of the suspicious tissue. Depending on the site, the biopsy is guided by interventional fluoroscopy or ultrasound imaging with a preceded CT or Magnetic Resonance Imaging (MRI) scan. The malignant tumors and their possible metastases can then be treated with chemotherapy, surgery or radiation therapy, for example.

Another increasingly prevalent medical condition is an abnormal rhythm of the heart, called arrhythmia. In the United Kingdom, for example, it is estimated to affect 2.35 % of the population [KCW+18]. It can usually be corrected with cardiac ablation, where a catheter is inserted through a blood vessel into the patient's heart and small scars are created using heat, i.e. radiofrequency energy, or extreme cold, i.e. cryoablation, which hinder some of the electrical signals which cause the irregular rhythm of the heart. For the correct positioning of the catheter, the surgery is usually guided with interventional X-ray imaging [May22].

There is one thing that all the aforementioned treatments have in common: some X-ray technique is, at some point, always used for non-invasive imaging, be it for screening, diagnosis or surgery.

However, two problems arise immediately: (1) X-rays are harmful for both the patient and the surgeon and might lead to more severe health issues and (2) X-ray images (e.g. for fluoroscopy) are merely projections and lack the third dimension, which is creating difficulties especially when navigating interventional instruments like needles or catheters inside the body.

The second problem can be solved with Computed Tomography (CT), which uses multiple X-ray projections to compute cross-sectional images which include the missing third dimension. However, this inherently reinforces the first problem. There are several ways to reduce the dose, e.g. by reducing the current of the X-ray tube or by reducing the number of projections that are used for the calculation of the CT images. Both of these dose limiting ways introduce artifacts, however, in the form of noise (predominantly with low X-ray currents) or streaking artifacts (predominantly with a low number of projections). These artifacts exist because data essential for the CT calculations are missing. Therefore, it is necessary to re-introduce data that is needed for an appropriate CT reconstruction without exposing the patients and surgeons to further X-radiation. This is done with the help of *prior knowledge*.

In recent years, machine learning – as one type of artificial intelligence – has found its way into medical imaging. Two related subtypes, *Convolutional Neural Networks (CNNs)* and *deep learning*, are especially suitable for image processing and reconstruction problems since they are built after the human vision and are trainable, i.e. given enough pairs of input and output data, the algorithms attempt to "learn" the most essential features and can eventually be used to "predict", i.e. extrapolate or interpolate, outputs from unseen input data.

1.2 Prior Knowledge

Before giving a definition for *prior knowledge*, it is necessary to understand what *knowledge* is. Depending on in which area the term is used, it is defined and interpreted slightly differently. One of the more general definitions is given by the Cambridge Dictionary [Cam]:

Definition (Knowledge)**.** Understanding of or information about a subject that you get by experience or study, either known by one person or by people generally.

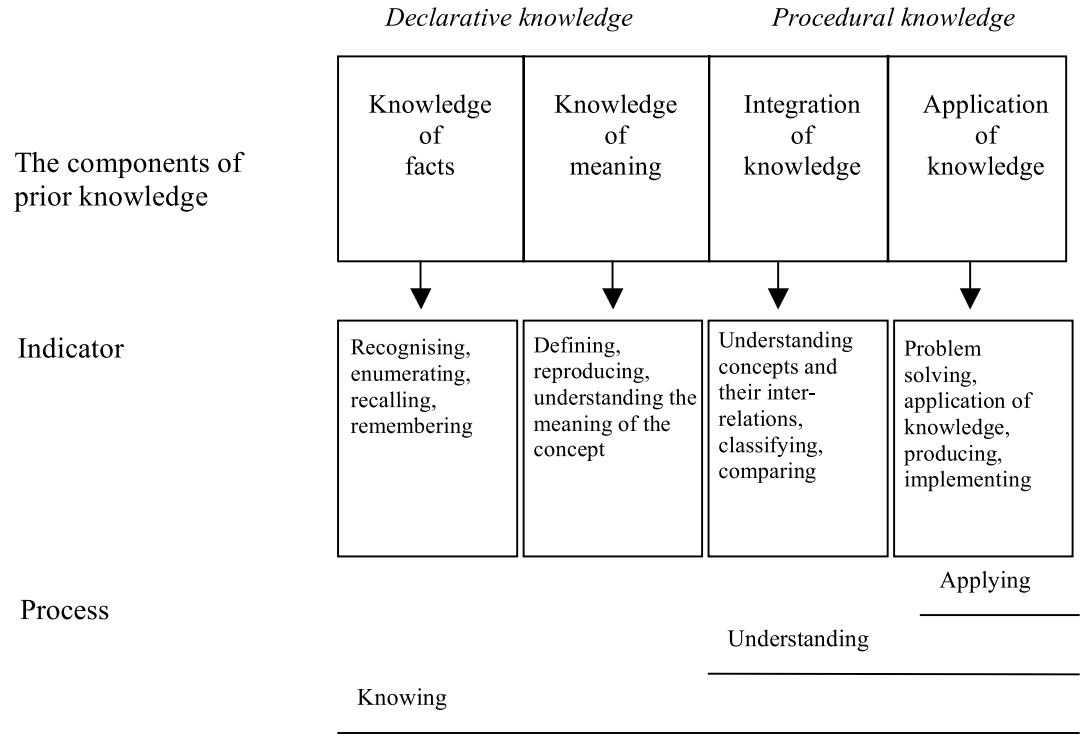


Figure 1.1: Model of prior knowledge described by Hailikari et al. [HNL07].

In the scope of this thesis, the term is closely related to the *scientific method*, which provides a systematic way of how to acquire knowledge scientifically: Based on a collection of data through observation, hypotheses can be formed and validated by experiments and analyses, which ultimately result in new knowledge [CS90].

Moving on to *prior* knowledge, it can now be defined as:

Definition (Prior Knowledge)**.** Knowledge which is available before (i.e. prior to) and potentially supportive in or necessary for solving a new problem using the scientific method.

A model of general prior knowledge and its components was described by Hailikari et al. [HNL07] and is shown in Fig. 1.1. More specifically, for the scope of this thesis, *prior knowledge* denotes every available useful information which can be used to optimize the reconstruction quality of artifact-bearing (interventional) CT images due to missing data, to reduce the exposure of patients and surgeons to X-radiation.

Since the term 'useful information' is rather fuzzy and is possibly interpreted differently and subjectively, the prior knowledge specific to this thesis will be divided into different categories:

- **Algebraic Prior Knowledge:** Knowledge about traditional, i.e. without machine learning, mathematical algorithms and methods for CT reconstruction, e.g. FBP, DBP, FDK or other iterative reconstruction algorithms.

- **Machine Learning Prior Knowledge:** Knowledge about data distributions that were learned/extracted from training artificial neural networks on specific data sets, but also about hyperparameters for defining the network architecture and to set appropriate optimization techniques.
- **Temporal and Model Prior Knowledge:** Knowledge about temporal changes and invariances in the image data, e.g. from planning scans before surgery or projections acquired moments before the current scan, or textures and shapes of specific tissues or materials.

It has to be noted that use of prior knowledge in the domain of scientific reasoning seems contrary to its purpose in this thesis: Much research has been put into if using prior knowledge should be minimized or even disregarded when reasoning scientifically. This may make sense for this purpose because prior knowledge limits the hypothesis and/or experiment space considerably, in a way that new hypotheses can only emerge from memory retrieval (and merely result in an incremental gain of knowledge), as opposed to building hypotheses solely based on observations (which might result in gaining disruptive knowledge) [KD88].

However, disregarding prior knowledge implies reasoning context-free. Though this might work for some research, it is not applicable for every domain. Therefore, creating hypotheses should be based on a good balance between prior knowledge and empirical observations [Zim00].

1.3 Objective and Research Questions

As the title of this thesis already suggests, the ultimate objective is the reduction of X-radiation exposure during CT-guided interventions for patients and surgeons while creating reconstructions with a quality similar to as if full dose was used.

As described in Sec. 1.1, there are two main types how to reduce X-radiation: (1) Low dose CT, where the energy of the X-ray beam is decreased such that fewer photons are emitted from the tube while the number of projection images is kept high. This way, estimation of the expected (attenuation coefficient) value of each detector pixel is undersampled, i.e., simply put, an increased noise level in the projections and consequently in the reconstructions. This type of undersampling is usually solved with statistical denoising methods. The focus of this thesis, however, is (2) sparse view CT, where the X-radiation is decreased by reducing the number of projections while keeping the radiation of each projection high. This way, the projections are close to noise-free, but they are undersampled themselves, resulting in locally non-invariant streaking artifacts in reconstructions. Though these artifacts are more difficult to suppress (due to the non-locality), algorithms designed for this type are not restricted to statistical denoising methods, leaving more space for creativity.

Having specified how the radiation is reduced, the question remains how the reconstructions can be enhanced to have a quality similar to full-dose scans. The previous sections have already given hints: Convolutional Neural Network (CNN) are going to be

used throughout the thesis, since they have state-of-the-art performance in many medical imaging-related tasks, and different kinds of prior knowledge will be incorporated into the methods to both limit the search space and guide the algorithms towards the correct solutions. To this end, the following research questions are attempted to be answered in the subsequent chapters:

Research Question 1 *How do the three different types of prior knowledge – Algebraic, Deep Learning and Temporal/Model Prior Knowledge – influence the quality of the final reconstructions?*
For each type, two methods will be presented in the scope of this thesis which primarily make use of the respective prior knowledge. Despite their fundamental differences in the computation of the final reconstructed images or volumes and their use cases, the setup of the data, i.e. the simulated acquisition setup, is bound to differ slightly, be it in the number of projections, the resolution of the detector, or the data set being used for the simulations, for instance. Nevertheless, it was attempted to keep these differences as small as possible to allow for a meaningful comparison in the end to appropriately answer this research question. As stated before, note that only the three mentioned types of prior knowledge are going to be investigated, which excludes other potentially helpful information from different prior knowledge types.

Research Question 2 *How well do different similarity metrics assess the quality of reconstructed images/volumes wrt. a specific task, and which metrics are most suitable for evaluating CT reconstruction quality?*
When it comes to assessing the quality of reconstructions, similarity metrics are the preferred way to quantify the performance. As it will turn out, some metrics are derived from others, and therefore merely change the scale, while other metrics are entirely different, mathematically. Showing a greater similarity using one metric does not necessarily imply the same correlation with another metric. For this reason, every presented method in the following chapters will not only rely on one metric, but the evaluation will always be carried out with several ones, such that it can be investigated which metric is suited best for a certain use case or CT image reconstruction in general.

Research Question 3 *How does the reduced X-ray exposure (by reducing the number of projections) in combination with the incorporated prior knowledge correlate with reconstruction quality and computation time, and what does this mean for medical applications?*
A major goal of this thesis is not only to explore which type of prior knowledge is most supportive for the reconstruction task, but also finding methods that could in fact be used in medical interventions. The best reconstruction algorithm is useless if it cannot be applied to data acquired during surgery because of a too high computation time or memory requirements that are only met with hardware dedicated for research or academia. Therefore, algorithms for supporting medical interventions should be a sufficiently good compromise between reconstruction quality, dose reduction and acqui-

sition/computation time, which is going to be investigated by this research question for the methods proposed in the following chapters.

1.4 Significance and Limitations

Though there are numerous methods and algorithms in the literature that make use of prior knowledge for the reconstruction of artifact-bearing CT images or volumes, none of them have given a definition of prior knowledge nor have they attempted to compare how the previously described types influence the quality of the reconstructions. In this thesis, a comparison of these types will be provided and discussed for the first time according to the best of the author's knowledge. This is hopefully going to make future research aware of why and how to classify the prior knowledge incorporated by novel methods. Despite being limited to CT in this thesis, the general concept of prior knowledge is well-known in every research area and can therefore be adapted or specialized differently to fit the demands of other scientific fields.

Ultimately, the main goal of this thesis application-wise is to reduce the X-ray dose that surgeons and patients are exposed to for imaging during operations or follow-up scans. This is very important in order to reduce the risk of getting acute or chronic radiation burns, or in the worst case developing cancer.

However, several assumptions are made in the scope of this thesis in order to concentrate mainly on the methodological part as compared to the practical applicability, i.e. for clinical daily routine, of the presented methods. Especially for the training of the neural networks, much data is necessary to cover most of the variance in the images. Though many *conventional* CT data sets are publicly available, *interventional* Cone Beam CT (CBCT) data sets are mostly absent. For this reason, the interventional scans that are used in this thesis are, unless stated otherwise, usually simulated based on scans from conventional systems. This means that errors and artifacts caused by real-world physics cannot be considered in many cases, e.g. beam hardening, motion artifacts, systematic positioning errors of the X-ray tube or detector due to the limited accuracy of the gantry or comparatively slow read-out time of the detector when the system is rotating.

When planning scans are incorporated as prior knowledge, they are assumed to be perfectly, or at least closely, registered to the interventional data. The inter-modal registration that would be necessary here can become – especially due to the presence of artifacts in the interventional scans – a very challenging task and is therefore not a focus of this thesis which essentially concentrates on the reconstruction process.

All methods that are described in this thesis make use of CNNs. Despite them already having millions of trainable parameters, several methods have been developed over the past years to visualize and attempt to understand how they work internally, making them more credible and explainable for future implementation and application on medical hardware, and thus less acting like a black box. Of course, other types of neural networks have existed before, like multilayer perceptrons, and new types, like generative adversarial networks or (vision) transformers, have been developed since. However, these

would need to have even more parameters trained (and thus need even larger amounts of data), e.g. hundreds of millions for just a single-layer perceptron or (vision) transformer for a 128x128 image, or have not yet been explored in the same detail as CNNs and are therefore less credible at this time, e.g. generative adversarial networks or (vision) transformers.

Although aiming for interventional CBCT reconstruction, some methods in the following chapters will be described for and evaluated on (conventional) parallel or fan beam CT. This will show that the general concepts of these methods can also be applied to different data. In most cases, these methods can also be modified to process cone beam projections, but this results – in some cases – in very high memory consumption and/or a slowdown of the reconstruction process, making the training of the respective Convolutional Neural Network (CNN) impossible, currently.

1.5 Thesis Structure

Chapter 1 has introduced and motivated the topic of this thesis. A definition for prior knowledge and its types has been provided and the research questions were stated. Moreover, the significance and limitations of this research were described and the structure of the thesis was summarized.

Chapter 2 will give an overview over the historical developments of both CT and deep learning to be able to understand which technologies are used for which use cases.

After the descriptions in Chapter 2, Chapter 3 is going to give more detailed explanations of the concepts of CT image formation and reconstruction mathematically and will define the most commonly used metrics and loss functions that will be used in the subsequent chapters.

In Chapter 4, related works and literature will be presented to put the thesis into scientific context. Furthermore, it will be pointed out which scientific gaps exist and which of these this thesis is trying to fill.

Based on the different types of prior knowledge defined above, Chapter 5 will provide at least one example for each type.

A rather unconventional chapter is going to follow afterwards. Chapter 6 will briefly present methods that failed or did not work out as expected. Not only will the unsatisfactory results be shown but it will also be attempted to give explanations and reasons for the failures.

Finally in Chapter 7, the previously described methods will be summarized in terms of their quality for real-world applications, and it will be discussed which benefits and downsides there are in the methods, and which kinds of prior knowledge should most likely be combined to achieve the best reconstruction results. The research questions will be answered explicitly, pointing out which scientific gaps have been filled, and a brief overview over what still needs to be explored will be given.

Chapter 2

Historical Developments

To get a better insight into why certain types of CT and artificial neural networks (or deep learning in general) are predominantly used for specific scopes of application, it is necessary to understand how they developed historically and what the original developments aimed for. Therefore, this chapter provides a brief summary of the most substantial developments and attempts to justify why only some of them will be mainly focused on in the remainder of this thesis.

2.1 Computed Tomography

Computed Tomography (CT) is an imaging technique based on X-ray measurements from different positions around the subject or object to be scanned. Compared to traditional X-ray images, CT allows producing cross-sectional images, called *slices*. Using appropriate techniques, not only two-dimensional images, but three-dimensional volumes can be computed from the acquisitions. The benefit of CT is that, unlike X-ray images, the resulting images are free of superimpositions [KS88, Chapter 1]. This has a wide range of applications in a medical context. It can, e.g., be used to detect infarction, tumors or calcification in the head, tumors or other changes in the lung, to detect coronary artery disease and investigate abdominal pain, to image complex fractures and to visualize vessels [LFM+15].

CT systems usually consist of three different types of components:

X-ray source In the scope of CT, the X-ray source is usually an X-ray tube which transforms electrical energy into X-rays. A high voltage is connected to the cathode and anode inside the vacuum tube, creating an electrical field where electrons are emitted from the cathode and collected by the anode. Due to the high energy of the electrons, they are able to convert the (loss of) kinetic energy to X-ray photons during the deceleration when being deflected by another charged particle in the anode, which is called *bremsstrahlung* [Cer16]. However, this only happens for about one percent of the electrons. The remaining energy is converted to heat. For this reason, rotating anode tubes

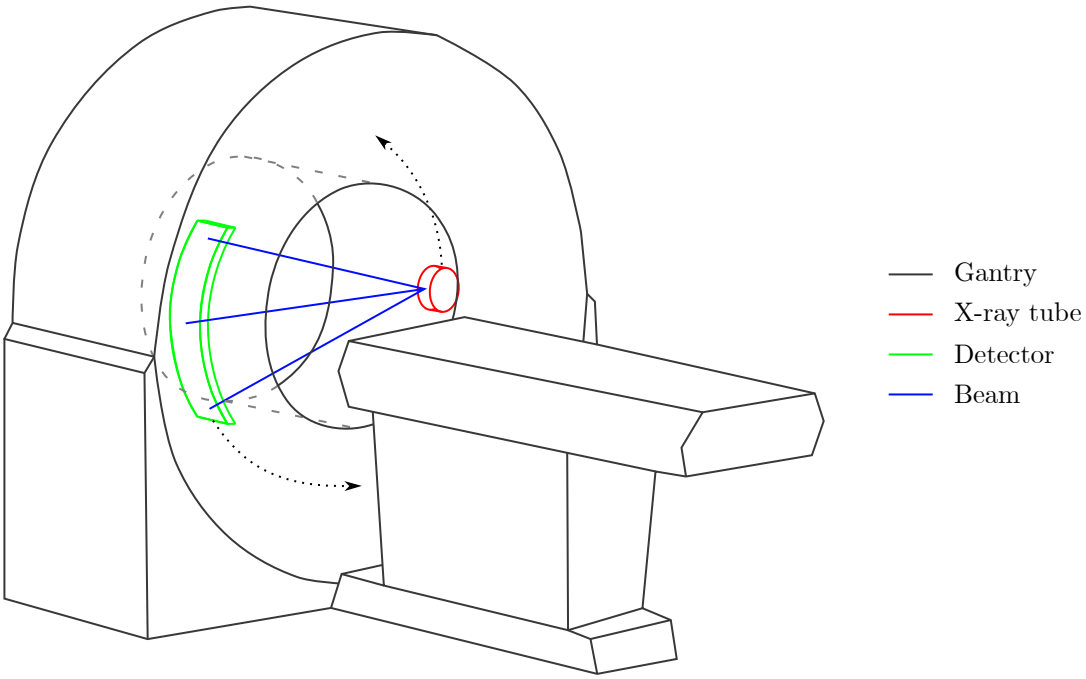


Figure 2.1: Schematic illustration of a conventional CT scanner. Arrows indicate the simultaneous rotation of the X-ray tube and detector.

were designed to distribute the heat on a larger area to reduce possible damage [Beh15, Sec. 1.3.6].

X-ray detector The X-ray detector is the component of the CT system that receives the photons emitted by the tube. Originally, sensitized glass photographic plates were used for this purpose, but they were quickly replaced by X-ray film containing crystal grains of light-sensitive (and therefore not only sensitive to X-rays) silver compounds. During exposure, the X-ray photons produce electrons in the film, which are trapped at defects of the crystals, eventually creating clusters of invisible atomic silver. The film is then developed with a chemical reaction which makes the clusters visible to the human eye [Mar06, Sec. 15.3.2]. When computers became powerful enough to store and process images, digital solid state X-ray detectors were developed and used for "live views" in angiography procedures or where many projections were needed in a short time, like CT. These digital detectors are able to directly convert X-ray photons to electrical charge, which can be read out and removed, i.e. reset, within a very short time [CDR99]. However, the image quality differs between digital and film radiographs. For this reason, film detectors have not yet been completely replaced by digital detectors. Moreover, there are other types of X-ray detectors, like dosimeters or Geiger counters, that are used for dose measurement, which is why they do not have an application for the imaging in CT systems and are not described any further here.

Gantry The gantry is the element that combines the other components of the system and holds them in place. For conventional CT scanners (see Fig. 2.1), it is usually shaped like a ring or doughnut. The X-ray tube and detector are mounted on opposite sides, i.e. fixed 180° apart, and the gantry allows them to rotate rapidly along the ring. The patient to be scanned is lying on a table through the hole of the gantry. For positioning purposes or acquiring three-dimensional scans, either the table can be moved through the gantry, or the gantry itself can move along the patient table. The gantries of interventional scanners are different: to provide more flexibility and space for the surgeons, the X-ray components are usually mounted on an arm that is shaped like the letter 'C', hence the name *C-arm CT*. This allows for more dimensions of freedom – the C-arm can be rotated about and translated along all three spatial dimensions to a certain degree – but on the other hand also limits the acquisition speed. More recent interventional systems comprise two C-arms that can move independently or together. These so-called biplane scanners do not only allow for faster acquisitions or two X-ray images at the same time but also make it possible to apply algorithms that make use of different tube energies for being able to discriminate tissues in the reconstructions more reliably. Gantries of systems used for radiotherapy are very similar to interventional C-arm CT scanners but replace the X-ray tube with a linear accelerator to create high-energy X-rays, and do not need a detector.

Moreover, additional components like collimators and filters are often used to form the beam and avoid certain types of artifacts [BB11].

Five different generations of CT devices have evolved historically, though only one of them is mainly used nowadays [Buz11]:

Beginning in 1968, the prototype of the first generation of CT scanners used a translation-rotation approach with *parallel* beams: an X-ray tube and a (single-pixel) detector, kept at a fixed distance with the subject to be scanned in between, were both translated at the same time to sample one line of projections. Afterwards, the tube and detector were rotated about the center and the next projection could be sampled. This process was very slow (up to 9 hours for a complete acquisition of a brain, not including the reconstruction) and inaccurate due to the need of mechanically translating and rotating the device during the acquisition. In 1971, Godfrey N. Hounsfield developed the first commercially available head CT scanner using the principles of this first generation, 'EMI Mark I', which reduced the acquisition time per axial slice down to only 5 minutes [Hou73].

The second generation was similar to the first one, but now, multiple beams were emitted from the X-ray tube at multiple detectors, i.e. a detector consisting of multiple detector pixels in a row. Due to the shape of the beams, this type of projections is called *fan beam* projections. This allowed for larger rotational steps between projections for further reducing the acquisition time per projection while increasing the number of detector pixels for a higher resolution of the final reconstruction. Again, G. Hounsfield developed a head CT scanner of this generation in 1974, 'EMI 5000', which now took 18 seconds to acquire the projections of a slice [RPI77].

Still, the X-ray tube and detector had to be rotated/translated and stopped to acquire a projection, which limits both the speed of a complete acquisition, due to the time that is needed for accelerating and decelerating the tube and detector for the rotation between the projections, as well as the accuracy of the angle that a projection is acquired at. This changed with the third generation, where the devices performed a continuous rotation during the acquisition. The CT systems of this generation, available since 1975, were the first ones to acquire scans of not only head but also chest and abdomen areas, because of an increased number of detector pixels (and thus larger fan angles of up to 60°) and a fast acquisition time per slice of about 20 seconds. This generation is the one that is mainly used for conventional CT systems to date, however with several improvements [Buz11].

For the sake of completeness, the remaining two generations will also be described here briefly, although they did not prevail against the third generation.

The CT scanners of the fourth generation were developed in 1977. The detector pixels were arranged in a full 360° circle around the subject, such that only the X-ray tube had to be rotated. The acquisition time per slice could be reduced to 1 to 5 seconds. However, this static arrangement of detector pixels corresponds to a multitude of rotating detectors that were used in the third generation, which quickly made this fourth generation obsolete.

The final, fifth generation used a completely static arrangement of both the tube and the detector. Like in the previous generation, the detector consisted of a full circle of detector pixels around the subject. The tube was larger as well: circular segments of tungsten anodes were placed next to the detector ring and X-radiation at different positions along these segments were created by deflection of electron beams of cathode-ray tubes. For this reason, the scanners of this generation are also called *electron beam* CT systems. Since the deflection can be changed very fast, the acquisition time could be reduced to 30 ms, which facilitated real-time imaging of the heart. Not only due to high costs and the technically more challenging construction of these systems did this generation become obsolete, but also because modern scanners of the third generation are able to produce reconstructions of similar time and spatial resolution at a much lower cost [MR06].

A further advancement in CT development was *helical* CT, which is based on the third generation but additionally moves the patient table during the acquisition resulting in a trajectory of a helical shape. This facilitates an even faster generation of truly three-dimensional reconstructions with isotropic voxels, compared to stacked axial slices, which were originally reconstructed with third generation systems [KSK+90].

Moreover, X-ray detectors were expanded to have multiple rows of pixels to further reduce the acquisition time and the slice thickness. Again, for this, now three-dimensional shape of beams, systems using this kind of detectors are named *cone beam* CTs and were introduced in 1998 [MPT+98]. Dual-source CTs were introduced, which consist of two X-ray tubes that are 90° apart and thus reduce the acquisition time by half and can also be run at different voltages to allow for a better separation of tissues and contrast agents in the reconstructions [FMB+06].

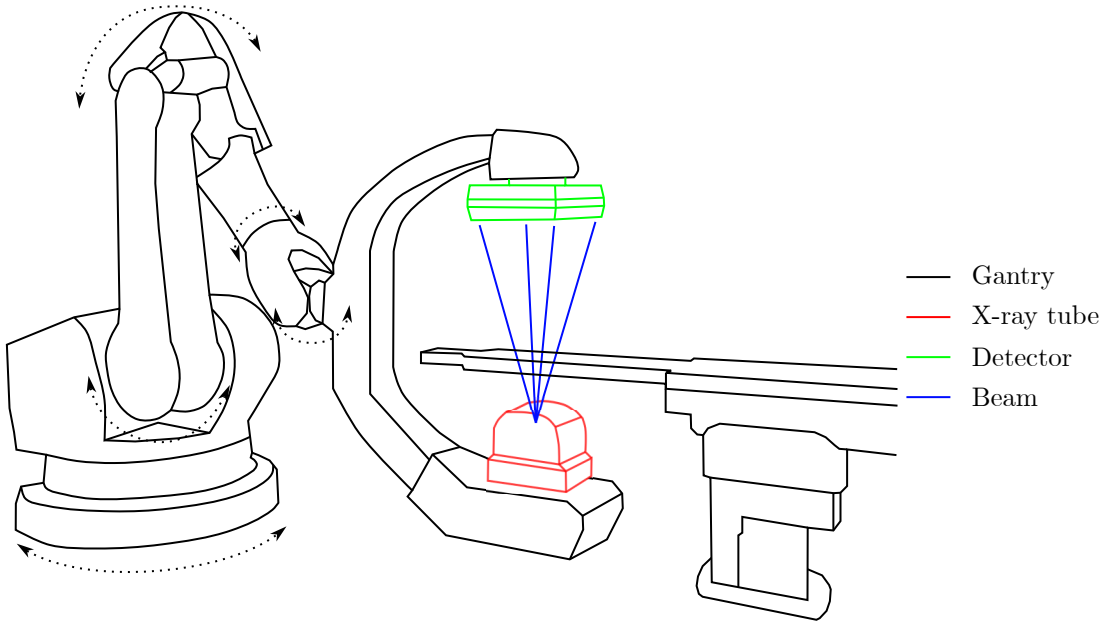


Figure 2.2: Schematic illustration of an interventional C-arm CT scanner. Arrows indicate the degrees of freedom to move the C-arm.

The main focus of this thesis, however, is not conventional (stationary), but *interventional* (C-arm) CT (iCT). Compared to conventional systems, CTs used during interventions usually do not have a closed tubular shape but consist of rather open arms, shaped like the letter 'C', with an X-ray tube mounted on one side and a flat panel detector, i.e. a flat multi-row detector, on the opposite side (see Fig. 2.2). This configuration can be rotated, translated and tilted around the patient table. Currently, these interventional systems are mainly used for fluoroscopy, i.e. real-time X-ray imaging, as a guidance tool for surgeons to track medical instruments. Truly three-dimensional Computed Tomography is still rather uncommon because of the relatively low rotation speed of the C-arms (compared to a third generation conventional CT system) as well as a higher amount of artifacts and inconsistent reconstruction of attenuation values with respect to conventional CT reconstructions [OWK09].

Mathematical explanations of the aforementioned CT generations and possible algorithms for reconstruction will be described in Chapter 3.

Comparing CT to other imaging modalities like MRI, CT has fewer restrictions for the scan, e.g. objects containing metal, like pacemakers, are allowed in CT, a certain intensity in the reconstruction should theoretically have a similar meaning, and the scan itself is performed very quickly. The biggest downside of CT is its use of X-radiation which can be harmful for human beings if a high dose is applied and especially increases the risk of developing cancer. For this reason, it is necessary to keep the exposure during a scan as low as possible.

There are two possibilities to achieve a low radiation dose. First, the X-ray tube

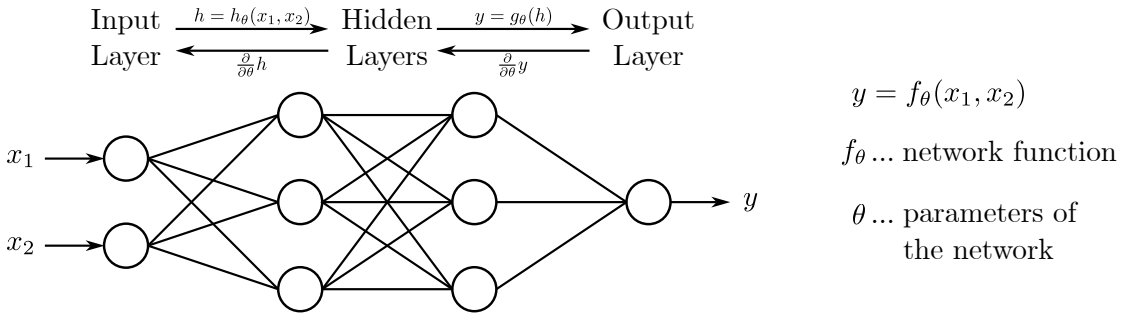


Figure 2.3: Simplified illustration of a multi-layer perceptron with two inputs x_1 and x_2, two hidden layers and one output y. Inference from left to right. Gradient descent based on partial derivatives calculated from right to left.

current exposure time product per acquisition can be reduced, which results in noisier projections and therefore noisier reconstructions. Second, the number of views, i.e. the number of projections from different positions, can be reduced while maintaining a high current time product, which mainly introduces streak artifacts due to the undersampling.

On the other hand, in the context of iCT, where it is mainly necessary to track the position of instruments over time, it is usually not feasible to perform a full scan and only few projections are acquired, which leads to a very poor reconstruction quality using the standard techniques, even if the scene may have changed only very little.

2.2 Artificial Neural Networks and Deep Learning

In the last years, machine learning and especially artificial neural networks and deep learning techniques have evolved drastically and also found their way to medical imaging and image reconstruction [ZD20]. For this reason, a brief summary of the developments, which are essential for the scope this thesis, is given in this section.

The first theoretical foundations for artificial neural networks were described by Warren McCulloch and Walter Pitts in 1943 [MP43], who derived a simple mathematical model as an abstraction of biological neural networks comprised of connected neurons to form a graph.

15 years later, in 1958, the *perceptron* was invented by Frank Rosenblatt [Ros57], which describes a neuron mathematically as a linear binary classifier, i.e. given a real-valued vector as input and an equally-sized vector of so-called *weights*, the perceptron outputs one out of two classes by applying a threshold as *activation function* to the scalar product between the input and weights. Although the number of components in these vectors is unlimited, Marvin Minsky and Seymour Papert showed in 1969 that it is not possible for a single perceptron to solve every problem, in particular linearly nonseparable functions like the Boolean XOR [MP69, Chapter 12].

However, multiple (parallelly, but especially sequentially) connected perceptrons (called *multi-layer perceptrons*, see Fig. 2.3) were assumed (and later shown) to be able to

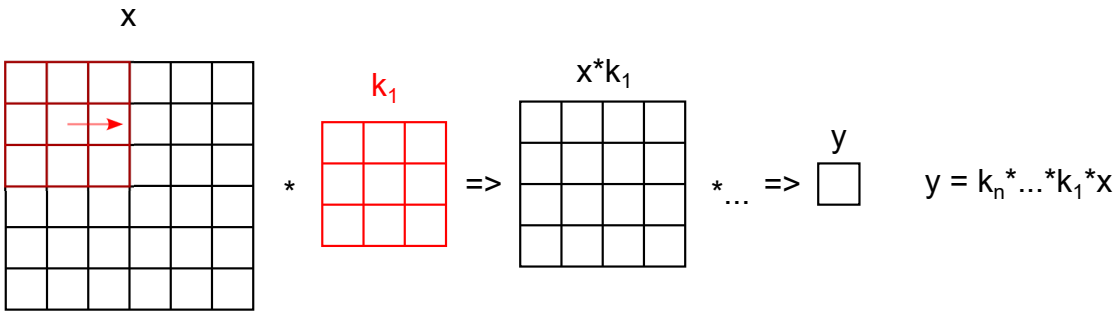


Figure 2.4: Simplified illustration of a Convolutional Neural Network for a two-dimensional image x with n convolutional layers outputting a single value y. The kernel of the first convolutional layer k_1 has a size of 3×3 and therefore 9 trainable weights. Since no padding is performed, the image shrinks by two pixel in each dimension after each convolution.

approximate any continuous function when replacing the thresholding by some non-affine activation function (*Universal Approximation Theorem*) [HSW89]. Using this property, Paul John Werbos was the first one to practically train multi-layer perceptrons in 1982 [Wer82] with the efficient error *backpropagation* algorithm, which was first described by Seppo Linnainmaa in 1970 [Lin70], i.e. iteratively (locally) optimizing weight vectors given corresponding input and output data, the so-called *supervised learning*, by distributing errors through the network while exploiting the differentiable properties of the single perceptrons, which became the de facto standard optimization technique to date.

Originally implemented for classification tasks, artificial neural networks were soon adapted to solve other types of problems, like natural language processing, medical image analysis and image restoration, due to progressively faster and more efficient hardware and algorithms, which allowed for networks with numerous neurons and layers, so-called *deep neural networks*. Moreover, semi-supervised and unsupervised algorithms have been proposed, as well as different kinds of architectures. The term *deep learning* was introduced to combine all of these approaches [LBH15].

One type of deep neural networks, which is especially important for medical imaging problems, is CNNs: Kunihiko Fukushima proposed the Neocognitron in 1980 [Fuk80], a multi-layer perceptron for pattern recognition tasks, which exploited the fact that most neurons only depend on a local neighborhood of pixels (the receptive field) in the input image, i.e. they have nonzero weights only for this neighborhood. These few weights are usually spatially invariant and, consequently, can be shared between all pixels, see Fig. 2.4. Mathematically, this is the concept of discrete convolution, hence the name Convolutional Neural Network (CNN). This way, the number of trainable weights can be reduced drastically (to the number of pixels in the local neighborhood, called the *kernel*) while retaining most of the capabilities of ordinary multi-layer perceptrons (in case of image-based problems). Since discrete convolutions can also be expressed in terms of weighted summations such like layers of perceptrons (as cyclic convolution matrices), CNNs are usually optimized using backpropagation, as well [LBD+89].

Connecting neural networks with CT, both the process of projecting image data and reconstructing image data from projections can be interpreted as an artificial neural network, at least in the discrete case using the CT system matrix (see Chapter 3). Because of this, it seems likely that neural networks can be used not only for reconstruction, e.g. one of the most widely used and direct algorithm called Filtered Backprojection (FBP) for reconstructions of single planes, i.e. slices, or the slightly modified Feldkamp-Davis-Kress (FDK) algorithm for three-dimensional reconstructions, i.e. volumes, but also for improving the image quality, since neural networks can be trained on data with optimal quality, which can be seen as a kind of prior knowledge for the reconstruction task.

However, in practice, this is not easily implementable due to different reasons: The number of trainable parameters for only the network that learns a full FBP is (approximately and without further assumptions) the product of pixels in the projection and in the reconstruction. Even for small images, this can lead to several million parameters and entails the need for a vast number of training samples as well as a long training time. Luckily, since the projection and FBP are mathematically well understood, most parameters can be set to a fixed/pre-computed value [WHC+18]. Still, there remain many trainable parameters that are responsible for reducing/removing artifacts caused by physical phenomena or undersampling. Simply training these using fully-sampled data can give pleasant visual results but lacks explainability, since the network artificially creates new data which may not necessarily be close to the true data. For this reason, there are still a lot of open questions in the field of deep-learning-based reconstruction of CT data that need to be investigated. Once solved, surgeons will benefit from high quality reconstructions and precise localizations of interventional instruments in a short time during an operation while patients are exposed to less X-radiation, which not only reduces the risk of developing cancer but also allows CT-guided follow-up surgery without much higher risks, if necessary.

Chapter 3

CT Image Formation and Mathematical Foundations

This chapter gives a brief mathematical introduction to the formation of X-ray projections, how the projected values can be reconstructed, i.e. how the process of projecting values can be inverted, and which problems arise in real-world applications, where the measurements are only available as discrete values.

3.1 X-Ray Measurements and Projections

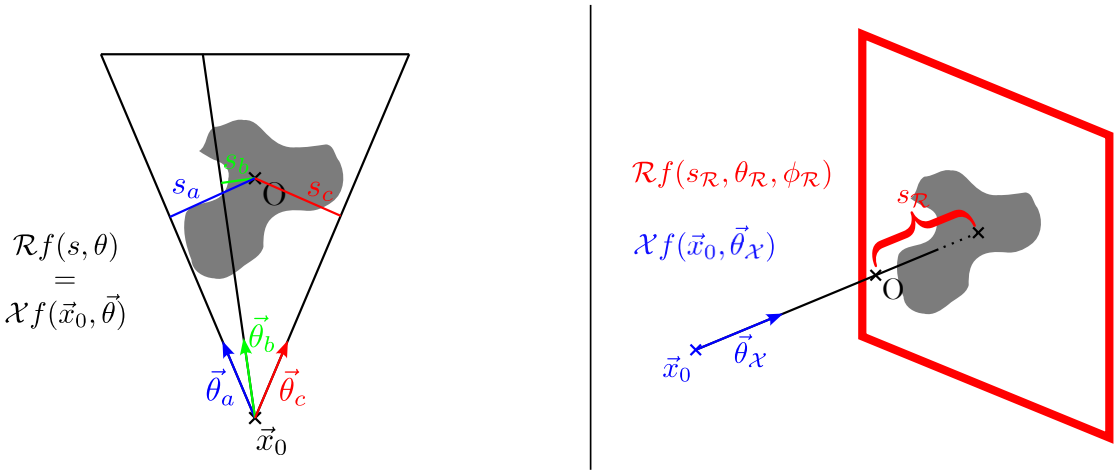


Figure 3.1: X-ray transform vs. Radon transform. Left: 2d, where the X-ray transform coincides with the Radon transform. Right: 3d, where the X-ray transform is the integral along the ray starting at $\vec{x}_0$ in the direction of $\vec{\theta}_\mathcal{X}$ and the Radon transform is the integral over the plane described by the distance $s_\mathcal{R}$ and the two angles $\theta_\mathcal{R}$, $\phi_\mathcal{R}$.

The physical process of measuring X-ray projections can be described in a very

simplified manner like this: The X-ray tube emits photons with a certain energy. In an idealized setting, these photons move along a ray from the tube towards the detector, which counts the number of incoming photons. Traveling to the detector, the photons pass through the object or subject that is located in between the tube and the detector. Due to its physical properties, the object or subject attenuates the (energy of the) photon which, in turn, reduces the probability of the photon to reach the detector.

Mathematically, this is described using the *Radon transform* or the *X-ray transform*, both of which coincide for the two-dimensional case, see Fig. 3.1.

3.1.1 Radon Transform

Consider a function of attenuation coefficients at every point in the plane $f : \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$. If for $f(x, y)$, it holds:

1. $f(x, y)$ is continuous,
2. the double integral
$$\int_{\mathbb{R}} \int_{\mathbb{R}} \frac{|f(x, y)|}{\sqrt{x^2 + y^2}} \, \mathrm{d}x \, \mathrm{d}y$$
converges,
3. for any point in the plane $(x, y) \in \mathbb{R}^2$:
$$\lim_{r \to \infty} \int_0^{2\pi} f(x + r \cos\theta, y + r \sin\theta) \, \mathrm{d}\theta = 0 \ \forall r \in \mathbb{R}_{\geq 0},$$

then the (two-dimensional) Radon transform of the attenuation function $\mathcal{R}f : \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$ is defined as
$$\mathcal{R}f(s, \theta) := \int_{\mathbb{R}} f(s \cos\theta - t \sin\theta, s \sin\theta + t \cos\theta) \, \mathrm{d}t$$
for any line in the plane defined by its (signed) distance to the origin s and its angle θ [Rad17].

In higher dimensions, the (generalized) Radon transform integrates over hyperplanes, which is not necessarily directly related to X-ray projections anymore. For this reason and for the sake of brevity, a formal definition is not shown here.

3.1.2 X-Ray Transform

Similar to the Radon transform, the X-ray transform calculates integrals of an attenuation function $f : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$, $n \in \mathbb{N}$, which is required to be continuous and with compact support. The difference to the Radon transform is that the X-ray transform calculates integrals of lines (instead of hyperplanes) in every dimensionality.

These lines are parameterized with a point $x_0 \in \mathbb{R}^n$ on the line and a directional vector $\theta \in S^{n-1}$ in the $(n-1)$-sphere, such that the X-ray transform $\mathcal{X}f : \mathbb{R}^n \times S^{n-1} \to$

$\mathbb{R}_{\geq 0}$ is defined as [Joh38]:

$$\mathcal{X}f(x_0, \theta) := \int_{\mathbb{R}} f(x_0 + t\theta)\, \mathrm{d}t.$$

This notation is much more suitable to describe the X-ray projections of the various beam types, from parallel over fan beams in the two-dimensional case to cone beam for projections of three-dimensional objects. For this reason, the X-ray transform will mainly used in the course of this thesis.

3.2 Backprojections and Direct Reconstructions

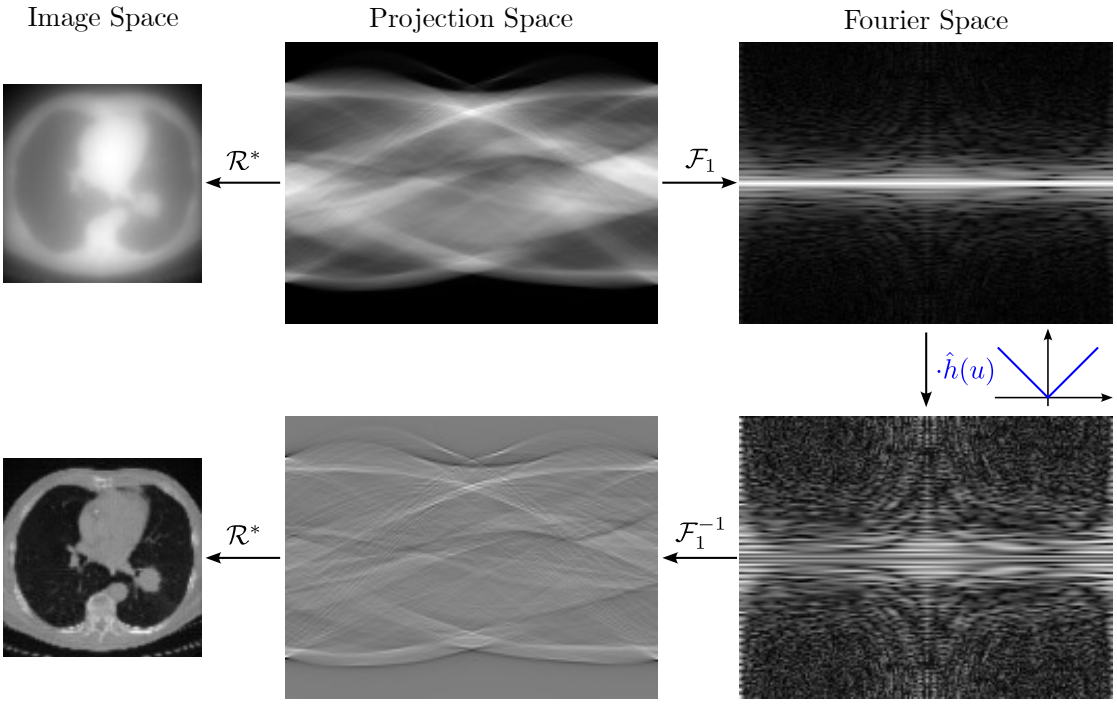


Figure 3.2: Steps of Filtered Backprojection for a parallel beam sinogram (top center). Filtering with a ramp filter is performed in Fourier space.

Some main goals of CT imaging are identifying or locating pathologies non-invasively or to trace instruments during surgical interventions without perspective distortions and superimpositions that exist in simple X-ray imaging, like fluoroscopy.

The previous section described how the projections of a function of attenuation coefficients is defined mathematically. To achieve the aforementioned goals, however, the *inverse problem* must be solved, i.e. instead of providing a mathematical definition for creating projections from a function of attenuation coefficients in space, algorithms for finding the attenuation function from already measured projections have to be derived.

The simplest idea is to "smear" back the values of the projections along the lines that they were defined for. This is called *backprojection* and is, mathematically, the adjoint Radon transform:

$$\mathcal{R}^* g(x,y) = \frac{1}{2\pi} \int_{\theta=0}^{2\pi} g(x\cos\theta - y\sin\theta, \theta)\ \mathrm{d}\theta$$

for the Radon transform $g := \mathcal{R}f$ of a function f satisfying the requirements in Sec. 3.1.1. In other words, the backprojection at a certain point is the integral of the projection values of all the lines that coincide with that point.

However, the backprojection g is not the original object function f (see Fig. 3.2 top left). It can be shown that the projections need to be convolved with a ramp filter (or modified versions of it) before backprojection to get back the original object function. This is called Filtered Backprojection (FBP):

$$f(x,y) = \frac{1}{2\pi} \int_{\theta=0}^{2\pi} (g(\cdot,\theta) * h)(x\cos\theta - y\sin\theta)\ \mathrm{d}\theta$$

with the one-dimensional filtering function h satisfying $\hat{h}(k) = |k|$ (the hat denoting the Fourier transform) as shown in Fig. 3.2.

As described in Sec. 3.1.2, the Radon transform is mainly used to describe parallel beam geometries, since this is what a regular sampling of its arguments results in. Following from this, the FBP algorithm was derived for two-dimensional parallel beams, as well. More commonly used geometries, like fan beam and cone beam (see Sec. 2.1), cannot be treated equivalently and need to incorporate modifications to the original FBP. The most straightforward method is to preprocess the measured projections by rebinning them to parallel beams and then applying the standard FBP. This rebinning step, however, is rather inaccurate and usually needs some kind of interpolation.

The FBP itself can be modified to avoid the additional rebinning: In case of fan beam geometries, the projections are multiplied with a cosine weighting (and possibly a redundancy weighting function to accommodate for redundantly acquired rays) before they are filtered and backprojected including a distance weighting [KS88, Ch. 3, Eq. 118-120].

Very similarly, the FBP was modified to handle cone beam geometries. Its three steps consist of multiplying the projection data with a weighting function, convolving them with a filtering function and backprojecting them onto the reconstruction grid [FDK84]. This algorithm is usually referred to as the FDK algorithm.

3.3 Differentiated Backprojection

Another possibility to reconstruct the function of attenuation coefficients for the specific case of cone beam projections from a circular trajectory was described by Dennerlein et al. [DNS+08]. The measured cone beam projections are treated as X-ray transforms $\mathcal{X}f(\mathbf{x}_0, \theta)$ from source locations $\mathbf{x}_0$ along a circular trajectory of radius R

$$\mathbf{x}_0(\lambda) = R \cdot (\cos\lambda, \sin\lambda, 0)^T, \quad \lambda \in \mathbb{R} \wedge R \in \mathbb{R}_+, \tag{3.1}$$

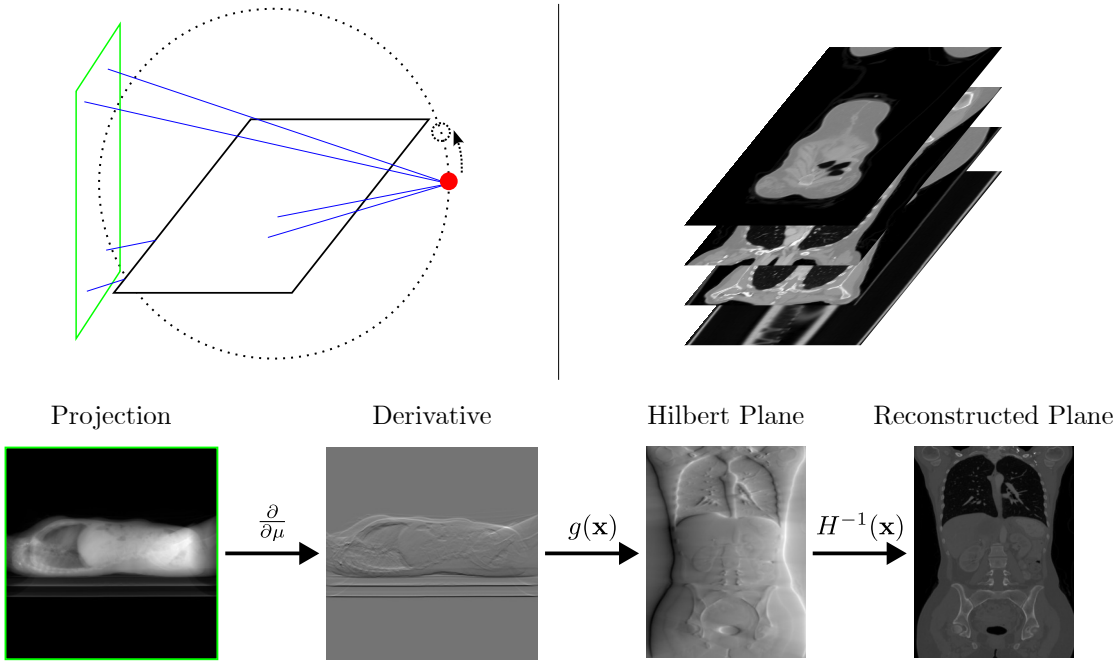


Figure 3.3: Top left: Acquisition setup. Top right: Stack of coronal planes of interest to be reconstructed. Bottom: Input (projections), intermediate steps (partial derivative, Hilbert plane) and output of DBP reconstruction algorithm.

along lines of direction $\theta \in S^2 \subset \mathbb{R}^3$ towards the detector, such that

$$\mathcal{X} f(\mathbf{x}_0(\lambda), \theta) = \int_{\mathbb{R}} f(\mathbf{x}_0(\lambda) + t\theta)\, \mathrm{d}t. \tag{3.2}$$

Applying the partial derivative along the source trajectory and backprojecting between the source locations $\mathbf{x}_0(\lambda), \lambda \in [\lambda_-, \lambda_+]$ results in the DBP

$$g(\mathbf{x}) = \int_{\lambda_-}^{\lambda_+} \frac{1}{\|\mathbf{x} - \mathbf{x}_0(\lambda)\|} \frac{\partial}{\partial \mu} \mathcal{X} f(\mathbf{x}_0(\mu), \theta)\Big|_{\mu=\lambda} \mathrm{d}\lambda, \tag{3.3}$$

which is related to the object function $f(\mathbf{x})$ by the Hilbert transform (see [DNS+08, Eq. 8]):

$$f(t, z) = \pi \int_{\mathbb{R}} h_H(t - \tau) \left(\hat{f}(\tau, z_1(\tau)) + \hat{f}(\tau, z_2(\tau)) \right) \mathrm{d}\tau. \tag{3.4}$$

This can, e.g., be solved by deconvolution and results in reconstructed planes perpendicular to the source trajectory. A visual representation of the steps of the algorithm is depicted in Fig. 3.3.

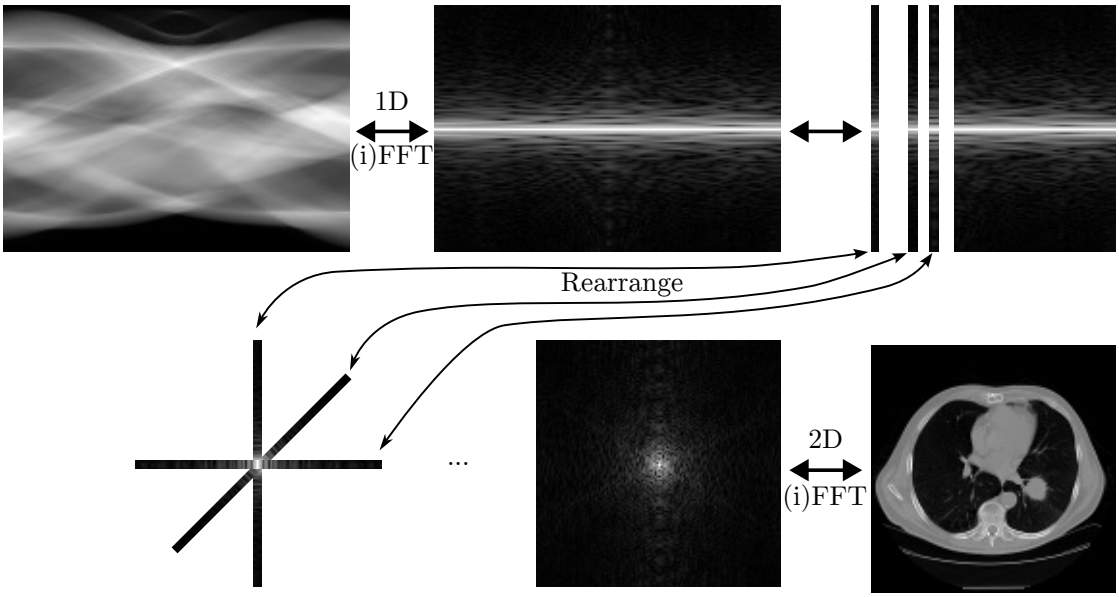


Figure 3.4: Illustration of Fourier slice theorem between sinogram/projections (top left) and reconstructed slice (bottom right).

3.4 Fourier Slice Theorem and Fourier Reconstructions

The Fourier slice theorem tells that the one-dimensional Fourier transform ($\mathcal{F}_1$) of a parallel-beam projection for a certain angle (p_θ) is the same as extracting a one-dimensional central slice with the same angle (S_θ) from the two-dimensional Fourier transform ($\mathcal{F}_2$) of the image function (f):

$$\mathcal{F}_1[p_\theta](u) = (S_\theta \circ \mathcal{F}_2[f])(u) \tag{3.5}$$

This means that every parallel-beam CT projection fills up the 2D Fourier space by one spoke, see Fig. 3.4. Though not directly applicable to the more commonly used fan- and cone-beam projections, the Fourier slice theorem has been used in reconstruction algorithms for these beam types with slight modifications [ZH95a; ZH95b].

3.5 Discretization

A major problem that has to be faced for real data is the discretization of the previously described equations for both projecting and reconstructing.

Assuming projections to be X-ray transforms $\mathcal{X}f : \mathbb{R}^n \times S^{n-1} \to \mathbb{R}_{\geq 0}$, the data is only known for discrete subsets of both dimensions. This is due to the physical properties of the detector, which comprises a countable finite number of pixels, and the gantry, which is usually limited to circular or helical trajectories, as well as the read-out of the detector, which can only be carried out at a countable finite number of gantry

configurations. Moreover, since the detector pixels only count the number of incoming photons, the image of $\mathcal{X}f$ is a discrete subset of $\mathbb{R}_{\geq 0}$ as well.

On the other hand, reconstructing images is usually performed on a compact regular Cartesian grid, i.e. the actual function of attenuation coefficients $f : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is sampled at a countable finite number of coordinates.

However, due to the discrete nature of both the measured projections and the reconstructed images, it is possible to express the process of projecting the attenuation function in terms of a simple matrix-vector product:

$$Af = p$$

with the so-called *CT system matrix* $A \in \mathbb{R}^{m \times n}$, the vector of attenuation coefficients $f \in \mathbb{R}^n$ and the vector of projection values $p \in \mathbb{R}^m$. With this notation, it is directly apparent that the reconstruction is an *ill-posed* problem: in general, the system matrix is not invertible, i.e. $\ker(A) \neq \{\mathbf{0}\}$.

From another point of view, the Fourier slice theorem tells that each parallel projection fills up the Fourier space by one line, which means that the number of projections for an exact reconstruction must be very high to accommodate for the larger gaps at the outer regions, i.e. high frequencies, of the sampled Fourier space.

Reducing the number of projections (which is one of the main goals of this thesis) has further mathematical implications. The fewer projections are available, the more elements exist in the nullspace $\ker(A)$, meaning the number of possible attenuation vectors, which created the projections, increases. Therefore, to reconstruct a credible image from few projections, the reconstruction algorithm has to regularize the set of attenuation vectors, which is mainly achieved by CNNs in this thesis.

3.6 Iterative Reconstruction

The direct reconstruction algorithms described in Sec. 3.2 and 3.3, or those based on the Fourier Slice theorem (Sec. 3.4) are most useful in deep-learning-based methods due to their fast computation time. However, they usually suffer from pronounced artifacts when the data is not fully sampled and, moreover, it is often not simple to incorporate further prior knowledge into these algorithms, e.g. about physical properties or shape information, i.e. *Temporal and Model Prior Knowledge*.

Based on the discrete representation of the Radon or X-ray transform, i.e. a linear equation system $Af = p$, several *iterative* reconstruction algorithms have been applied to or derived for CT data. Due to their inherent capability of incorporating Temporal and Model Prior Knowledge – and thus limiting the nullspace – the quality of the final reconstructions is often superior to direct methods at the expense of computation time and model complexity, rendering them hard to incorporate in deep learning methods. For this reason, they are only briefly described in this section, as the main focus in the next chapters will be on direct reconstruction algorithms.

Iterative reconstruction methods can be divided into two categories:

Algebraic reconstruction methods Based on algebraic properties of the linear equation system, this type attempts to solve for the reconstruction space variable with algorithms similar to (stochastic) gradient descent, Newton's method or quasi-Newton's methods. Examples include ART [GBH70], SIRT [Gil72] and SART [AK84].

Statistical reconstruction methods This type assumes the variables of the equation system to be samples of a random distribution, estimating the expected value of each pixel/voxel in the reconstruction, i.e. expectation maximization or maximum likelihood algorithms. This includes denoising methods in projection space, e.g. [LBV06], ordered-subsets algorithms, e.g. [EF99], and regularized methods, e.g. [KM75].

3.7 Physical Units

The two physical units that are most commonly associated with CT measurements are (mass) attenuation coefficients and Hounsfield units.

The linear attenuation coefficients describe how much the energy of an X-ray photon is attenuated at a specific point on the ray between the X-ray source and the detector. It is usually denoted with μ and is measured in cm^{-1}. It is defined as:

$$\mu = \frac{\Delta N}{N \Delta x}$$

where N describes the total number of photons along an X-ray, ΔN the number of photons removed and Δx the length of the section of the X-ray where the attenuation happens [McK98].

However, the linear attenuation coefficient does not solely depend on material properties, like the density and the atomic numbers, but also on the photon energy. To make it at least independent of the density, the mass attenuation coefficient is more commonly used, which is measured in $\mathrm{cm}^2\,\mathrm{g}^{-1}$ and is defined as:

$$\mu_M = \frac{\mu}{\rho}$$

where ρ denotes the density [McK98].

Since the values of the attenuation coefficients are rather small for tissues of the human body (e.g. $0.4\,\mathrm{cm}^{-1}$ for rather highly absorbing cortical bone tissue at an energy level of 100 keV) and because they still depend on the photon energy, Godfrey Hounsfield proposed a scale that is derived from the linear attenuation coefficient and describes how it changes relatively to water and air of similar X-ray properties [DCM+14, Sec. 11.2.2]:

$$\mathrm{HU}_{\mathrm{material}} = \frac{\mu_{\mathrm{material}} - \mu_{\mathrm{water}}}{\mu_{\mathrm{water}}} \cdot 1000$$

This, by definition, makes water have 0 HU and air (with an attenuation coefficient of $0\,\mathrm{cm}^{-1}$) have -1000 HU. It is important to note that the Hounsfield units in CBCT

scans do not exactly correlate to conventional CT scans and therefore should not be used to, e.g. estimate bone density/quality [MD07].

For evaluation and validation purposes of the methods presented in the following chapters, the Hounsfield scale is usually going to be used. For the optimization and training of the CNNs however, the data will usually be kept in attenuation coefficients to avoid negative values which might have unpleasant mathematical implications in the algorithms.

3.8 Metrics and Loss Functions

For the evaluation of the methods that are presented in the following chapters, it is necessary to define several error and similarity metrics. This allows for a quantitative comparison between not only the methods presented in this thesis, but also among those published in other articles. Although the metrics described in this section are widely used in the literature for regression problems, like image restoration and reconstruction, and give insights into how different methods compare, they can only show tendencies for the quality of rather abstract features: some metrics calculate a statistical value from pixel-wise differences or differences in a small neighborhood, others transform the image to a frequency domain before and another type computes the differences between features extracted from a CNN. None of these widely used metrics include information for task-specific evaluation, which is the reason why they should not be trusted blindly and a human expert study should always be conducted before the method is used in clinical practice. A tool for medical evaluation was developed in the course of this thesis and is described in Apx. C.

Some of the metrics that are described in this section, and sometimes combinations of them, are also commonly used as loss functions for the training of CNNs or neural networks in general. However, choosing a specific metric M_A as the loss function does not imply optimality wrt. itself during validation. As will be shown in the next chapters, optimal validation values for M_A might only be achieved when choosing a different metric M_B as the loss function. For this reason, relying on evaluating with only one metric can be misleading, which is why all the experiments in the following chapters will be evaluated using several metrics. This will also give a better general impression of the quality of the reconstructions (in accordance with the previous paragraph). Further information about loss functions in general can be found in Sec. 4.3. In fact, the loss functions described in this section are merely the data mismatch term Φ.

The images that the following metrics are defined on are defined as $F, G \in \mathbb{R}^{M \times N}$, unless stated otherwise.

3.8.1 Mean Squared Error

Due to its pleasant differential properties, the Mean Squared Error (MSE) (also called L2 Error) is often the first choice as an objective function for regression problems. The

MSE is defined as:

$$\mathrm{MSE}(F,G) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(f_{ij}-g_{ij})^2$$

Squaring the errors results in a very high importance for outliers. On the other hand, it is not easily recognizable what absolute pixel differences an MSE value stands for. For this reason, the Root Mean Squared Error (RMSE) was defined as:

$$\mathrm{RMSE}(F,G) = \sqrt{\mathrm{MSE}(F,G)},$$

to get an estimate of the absolute pixel differences, and the Normalized Mean Squared Error (NMSE), defined as [HSY20]:

$$\mathrm{NMSE}(F,G) = \frac{\mathrm{MSE}(F,G)}{\mathrm{MSE}(F,0)}, \quad 0 \in \mathbb{R}^{M\times N}$$

scales the value to the unit interval wrt. the target image F and therefore facilitates comparisons of results from differently scaled data (e.g. different data sets or imaging modalities). There are, however, slight differences in the definition of NMSE in the literature, which are derived from the differing definitions of the Normalizaed Root Mean Squared Error (NRMSE):

$$\mathrm{NRMSE}_1(F,G) = \frac{\mathrm{RMSE}(F,G)}{F_{max}-F_{min}}$$
$$\mathrm{NRMSE}_2(F,G) = \frac{\mathrm{RMSE}(F,G)}{\bar{F}}$$
$$\mathrm{NRMSE}_3(F,G) = \frac{\mathrm{RMSE}(F,G)}{Q_3-Q1},$$

F_{max} and F_{min} denoting the maximum and minimum value of F, $\bar{F}$ is the mean value of F [Zam89], and Q_1 and Q_3 denote the first and third quartile of the values in F.

3.8.2 Peak Signal-to-Noise Ratio

The Peak Signal-to-Noise Ratio (PSNR) is closely related to the MSE but evaluates the errors on a logarithmic scale, which is especially helpful for data with a high dynamic range. It is defined as:

$$\mathrm{PSNR}(F,G) = 20\cdot\log_{10}\left(\frac{I_{max}}{\sqrt{\mathrm{MSE}(F,G)}}\right), \quad I_{max}\in\mathbb{R}$$

with I_{max} being the maximum possible pixel value of the image.

This metric is not necessarily correlated to image quality as perceived by humans and should therefore always be evaluated in conjunction with other metrics [HG08]. Other problems of this metric include: the PSNR is undefined for $F = G$; I_{max} is not always sensibly defined (e.g. X-ray attenuation coefficients are unbounded), in which case some value must be set heuristically, which might render this metric incomparable to its use in other articles. The influence of this normalization factor on the metric is derived in Appendix B.2 and discussed in Sec. 7.1 as part of the answer for Research Question 2.

3.8.3 Mean Absolute Error

The Mean Absolute Error (MAE) (also called L1 Error) is favorable compared to the MSE when outliers should have a similar importance compared to expected values. The MAE is defined as:

$$\mathrm{MAE}(F, G) = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} |f_{ij} - g_{ij}|$$

3.8.4 Structural Similarity Index Measure

The Structural Similarity Image Measure (SSIM) is one of the perception-based metrics. Compared to the previously described metrics, the SSIM is not calculated from a per-pixel measure but from local neighborhoods. It is based on a weighted product of the *luminance* $l(F, G)$, *contrast* $c(F, G)$ and *structure* $s(F, G)$ [WBS+04]:

$$\mathrm{SSIM}(F, G) = l(F, G)^{\alpha} \cdot c(F, G)^{\beta} \cdot s(F, G)^{\gamma}, \quad \alpha, \beta, \gamma \in \mathbb{R}_{>0}$$

The luminance $l(F, G)$ compares the average intensity between the two images. The contrast $c(F, G)$ compares the standard deviation of the pixel intensities between the images. The structure $s(F, G)$ measures the joint variability of the pixels of the images, which, in case of medical images, provides information about the anatomical difference between the two images.

When used as a loss function for network optimization, it is usually modified to be

$$\mathrm{L}_{\mathrm{SSIM}}(F, G) = 1 - \mathrm{SSIM}(F, G)$$

to make the maximization of the SSIM a minimization problem.

Similar to PSNR, the functions $l(F, G)$, $c(F, G)$ and $s(F, G)$ include a normalization factor, setting the dynamic range of intensity values, to make the metric output sensible values (in fact, to stabilize divisions with weak denominators in these functions). Again, X-ray attenuation coefficients are unbounded, such that a sensible value for this factor needs to be set heuristically. The influence of this normalization factor on the metric is derived in Appendix B.3 and discussed in Sec. 7.1 as part of the answer for Research Question 2.

3.8.5 VGG Loss

The VGG Loss is a perception-based loss function that was introduced for the purpose of quantifying the similarity of the structural contents of images [LTH+17], as compared to pixel-wise losses/metrics. It is based on a trained VGG19 network [SZ15], which was originally presented for image classification. The VGG architecture consists of a number of convolutional layers for feature extraction and additional fully-connected layers, i.e. a multilayer perceptron, for the final classification, which makes 19 layers in total. For the VGG Loss, the weights from the network trained on ImageNet are used, the classification

layers are removed and the MSE between the feature maps, i.e. the output after the convolutional layers, of two images is calculated:

$$\mathrm{L}_{\mathrm{VGG}}(F, G) = \mathrm{MSE}(\phi(F), \phi(G))$$

where $\phi : \mathbb{R}^{m \times n} \to \mathbb{R}^{o \times p}$ describes the feature maps that are computed by feeding an image to the trained VGG network. Since the original network was trained to predict the classes of images, the feature maps are expected to contain all the information necessary for classification in the form of high-level features.

However, despite its likely useful properties, this loss function will not be used in this thesis and merely serves as an example for a loss function that is very close to the human perception: It was designed for natural images and photographs instead of medical images, so it does not necessarily work as well and might need to be retrained on medical data sets. Moreover, the VGG19 network consumes additional memory on the GPU which, in turn, reduces the amount of memory that can be used for the actual training. It also slows down the training process because it needs to perform many convolutions, especially when larger images are used.

3.8.6 Dice Loss

The Dice loss is derived from the Dice similarity coefficient:

$$\mathrm{Dice}(P, Q) = \frac{2|P \cap Q|}{|P| + |Q|}$$

for two sets P and Q. Although no assumptions are made regarding the elements of the two sets, the Dice similarity coefficient is usually used to measure the accuracy of segmentation problems, where one set contains the pixels/voxels of the segmentation output of an algorithm and the other set contains the pixels/voxels of the ground truth segmentation.

Segmentations of images are often given as implicit representations, i.e. matrices $F, G \in \{0, 1\}^{M \times N}$, instead of sets of pixels. For these types of segmentations, the Dice similarity coefficient can equivalently be expressed as

$$\mathrm{Dice}_{\mathrm{imp}}(F, G) = \frac{\epsilon + 2\sum_{i=1}^{M}\sum_{j=1}^{N} f_{ij} g_{ij}}{\epsilon + \sum_{i=1}^{M}\sum_{j=1}^{N} f_{ij}^2 + g_{ij}^2},$$

$\epsilon > 0 \in \mathbb{R}$ being a smoothing parameter pulling the value towards 1 and also avoiding division by zero in case of empty segmentations.

Since network optimization is usually performed on continuous variables both for the parameters and outputs (as opposed to the NP-hard integer programming), the segmentations are assumed to be real values as well, i.e. $F, G \in \mathbb{R}^{M \times N}$ like defined at the beginning, which calls for a fuzzy representation of the Dice similarity coefficient. Fortunately, the previously defined Dice on implicit segmentations can already be interpreted as a fuzzy version when relaxing the domains from $\{0, 1\}^{M \times N}$ to $\{x \in \mathbb{R} : 0 \leq x \leq 1\}^{M \times N}$,

as it already utilizes operations on real values and coincides with the Dice coefficient for binary values:

$$\mathrm{Dice}_{\mathrm{fuzzy}} := \mathrm{Dice}_{\mathrm{imp}}.$$

Assuming segmentations F, G of the relaxed domain, $\mathrm{Dice}_{\mathrm{fuzzy}}$ outputs values in this same domain, as well. The closer the value gets towards 1, the more similar the two segmentations are. Since target functions used for network optimization are usually designed to be minimized, the Dice loss function is defined as

$$\mathrm{L}_{\mathrm{Dice}}(F, G) = 1 - \mathrm{Dice}_{\mathrm{fuzzy}}(F, G),$$

which effectively inverts the fuzzy Dice coefficient.

An advantage of the Dice loss compared to other segmentation losses (e.g. binary or categorical cross-correlation) is that it inherently handles the size of the segmented areas/volumes by definition. For this reason, no extra weighting parameter needs to be set to balance the sizes of classes (e.g. foreground vs. background) which is necessary for an unbiased training of the networks.

The attentive reader might wonder why a segmentation loss is included in this section, whereas the topic of the thesis is image reconstruction. Semantic image segmentation is one of the most researched areas in medical image processing, which is why highly optimized algorithms and network architectures have been developed that are even able to achieve superhuman accuracy [LZL+17]. Exploiting this knowledge, one of the methods presented in Chapter 5 combines the reconstruction with an additional auxiliary segmentation task which ultimately helps the network to extract which information is beneficial for the actual reconstruction.

Chapter 4

State of the Art

In this chapter, several previously published methods and algorithms wrt. CT image reconstruction, deep learning and interventional CBCT will be described briefly, which can be considered the state of the art for their specific areas of application.

Since there is only a very limited number of publications specifically regarding deep learning-based interventional CBCT reconstruction, this chapter will not exclusively contain descriptions for methods of this specific problem but also more general areas of research.

For the purpose of a more systematic review of the methods, the following sections will be separated by the type of prior knowledge (see Sec. 1.2) that they are based on. It should also be noted that a strict classification of the methods into the types of prior knowledge is – in many cases – not possible. However, the type that is predominant is usually easily identifiable and will therefore be used for the classification here.

4.1 Algebraic Prior Knowledge

In the scope of this thesis, Algebraic Prior Knowledge describes algorithms not based on machine learning which are used for CT reconstruction. More generally, this includes all algorithms that can be used to solve inverse problems of different kinds. CT reconstruction, in general, is an ill-posed inverse problem, mathematically, and therefore can be solved more accurately if the reconstruction algorithm allows incorporating information on how to regularize the space of suitable solutions (see Sec. 3.5).

Despite their rather low reconstruction quality, especially for undersampled data, the FBP and, its three-dimensional counterpart, the FDK are still widely used in academic research as well as practically implemented on actual hardware used for clinical routine scans. The advantages of these direct reconstruction methods excel the usually better quality of the reconstruction methods that will be described in the following paragraph: high speed and low memory requirements (which is also particularly vital when training neural networks). It is possible that doctors prefer the lower quality of these algorithms because they are used to seeing the artifacts that are introduced by them, such that they build an intuition of what the data would probably look like without the artifacts.

The counterpart to these direct reconstruction methods are iterative ones (Sec. 3.6). These usually result in higher quality reconstructions but need considerably more time which makes them insufficient for interventional use. As previously discussed in Ch. 3, the two main types of iterative reconstruction algorithms can be categorized as algebraic and statistical reconstruction methods. Depending on the type and sparsity of the available data, one of the two types of algorithms is preferred over the other: Algebraic methods are often used when sufficient but artifact-bearing data (e.g. metal artifacts) is available, whereas statistical methods are often preferred when the sampling of the data is very low but can be represented well as a set of samples from a certain probability distribution.

For examples of the respective types of iterative reconstruction algorithms, the reader is referred to the descriptions in Sec. 3.6 to avoid redundant explanations here.

4.2 Machine Learning Prior Knowledge

Since this thesis investigates how CNNs can be used for (interventional) CT reconstruction, the methods described in this section will mainly focus on the domain of supervised deep learning. Nevertheless, there are other categories of machine learning, i.e. unsupervised or reinforcement learning, as well as other models, e.g. decision trees, support-vector machines or Bayesian networks. Only few of these other types have been applied to CT reconstruction problems, probably due to their inherent (or originally proposed) purpose for solving other problems, e.g. 'white box' decision trees for comprehensible and explainable classification (but relatively low accuracy and robustness), support-vector machines for linear or non-linear separation problems (and therefore merely binary classification), or Bayesian networks for probabilistic predictions of which inputs cause the presence of a certain output (but exact inference is NP-hard [Coo90], such that real-world problems can only be solved with broken-down or oversimplified versions like naïve Bayes networks [Zha04] or by introducing additional constraints on the probabilities).

The deep learning architecture that has gained the most attention for image processing tasks in general, and which is now also widely accepted to be used in algorithms in medical contexts including image reconstruction, is the UNet [RFB15b]. When it was originally presented in 2015, it was used for the segmentation of neuronal structures in electron microscopic stacks, achieving results that surpassed other segmentation methods by a large margin, not only in quality but also processing time. The UNet quickly found its way to all kinds of segmentation and eventually reconstruction and restoration problems in both medical/biological contexts and natural image processing. The success of this architecture lies in the skip-connections: Prior to the UNet, encoder-decoder architectures had been used for segmentation, already, but lacked the ability to reconstruct fine structures from the encoded features [LSD15]. Ronneberger et al. proposed to concatenate the feature maps of every encoding stage to the features of the corresponding decoding stages. This way, the high-resolution information is passed over to the decoder and, in turn, can be used more effectively for the segmentation task.

Milletari et al. [MNA16] replaced the two-dimensional convolutions with their three-

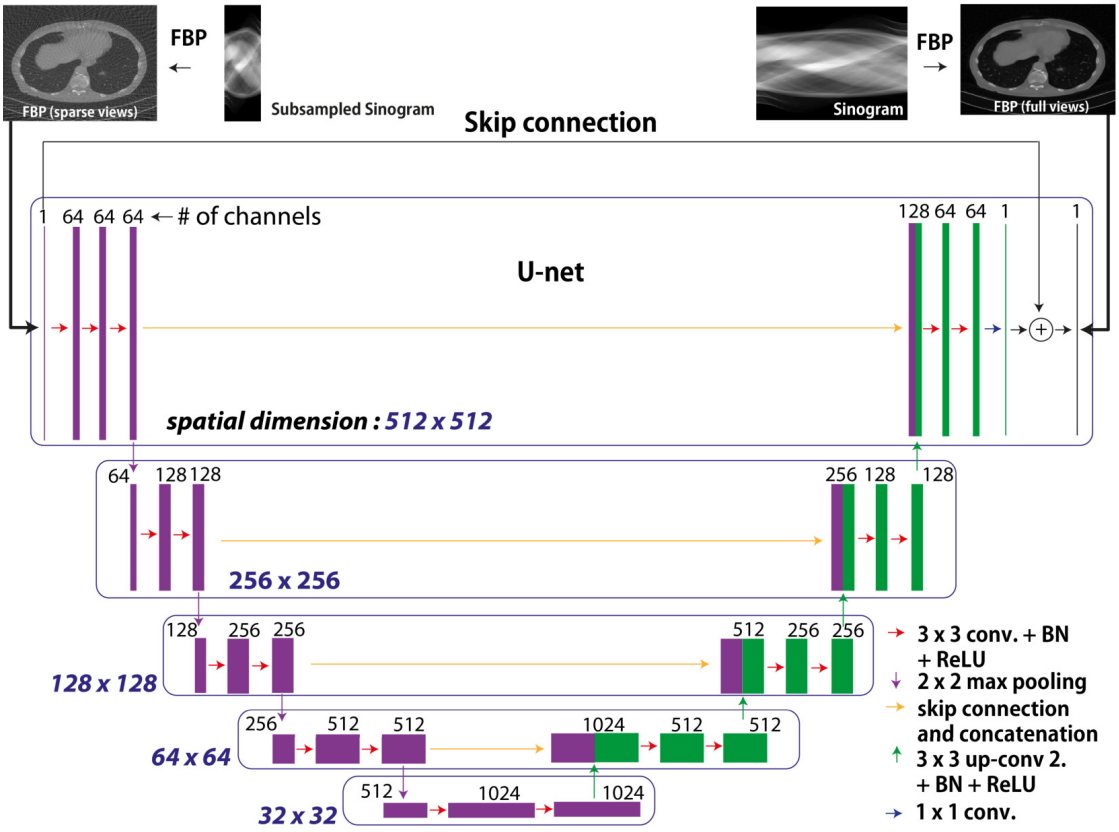


Figure 4.1: FBPConvNet [JMF+17] based on UNet [RFB15b] for sparse view CT reconstruction © 2017 IEEE.

dimensional counterparts and used residual blocks in every layer to enhance the gradient flow, which they called 'V-Net', in order to apply the network on volumes directly instead of stacks of images. More importantly, they proposed the Dice loss, a fuzzy version of the Dice coefficient, which can be used as an objective function for the network optimization directly. Compared to the previously usually used (weighted) crossentropy or logistic loss, the Dice loss is not prone to unbalanced classes (e.g. foreground vs. background) and can be applied to the segmentation of both large and small structures without any modification.

Both the UNet and the Dice loss are presently the go-to methods when it comes to choosing a CNN architecture and loss function for segmentation problems, just like the UNet with mean squared error loss for many kinds of regression problems (including image restoration and reconstruction).

In the scope of CT reconstruction, one main problem of merely applying a CNN like a UNet on the sinograms/projections (e.g. UDNN [BBP+19] or Sinogram Synthesis [LLK+19]) or reconstructions (e.g. FBPConvNet [JMF+17], see Fig. 4.1) is the absent data consistency wrt. the original raw data (e.g. not retaining properties of the

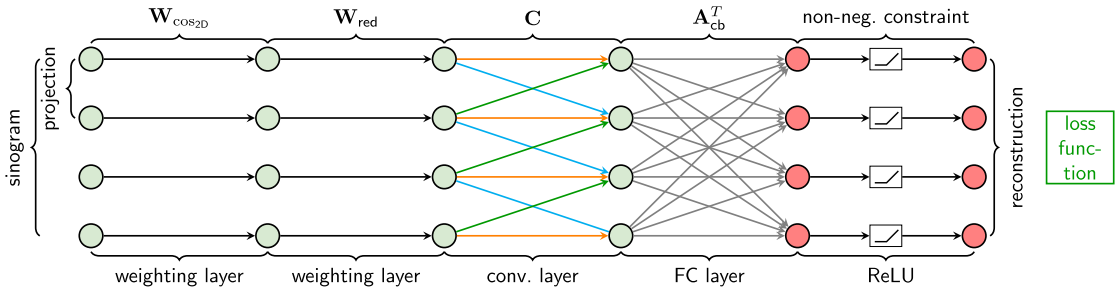


Figure 4.2: Precision learning for CBCT: implementing the backprojection as a neural network layer (A_{cb}^{T}) [WHC+18] © 2018 IEEE

Radon transform when applied on projections or getting different projection values when reprojecting "optimized" reconstructions). However, as previously described in Sec. 3.5, the Radon transform, and therefore the forward and (filtered) backprojection, can be approximated and expressed as matrix multiplications. This makes these transformations implementable as neural network layers, as proposed by Würfl et al. [WGC+16; WHC+18], which enables switching between projection and reconstruction space directly inside the network, see Fig. 4.2. Since these layers do not have optimizable parameters, they have an increased robustness and a decreased amount of data needed for training compared to fully trainable transformations like AUTOMAP [ZLC+18] (see Sec. 6.1).

Since iterative reconstruction algorithms usually result in a higher quality reconstructions compared to FBP, several works have focused on imitating these using neural networks. For this purpose, the iterations were 'unrolled' and each iteration comprised the necessary steps of the original algebraic algorithm, possibly including forward/backprojection layers, enhanced by trainable convolutional or other neural network layers.

One notable method is the LEARN network [CZC+18] for removing noise from low-dose CT acquisitions where in each iteration, the current estimation of the reconstruction is optimized with blocks of convolutional layers and data consistency wrt. the projections is encouraged by a weighted algebraic term corresponding to one step of gradient descent like in Algebraic Reconstruction Technique (ART).

Another method of this unrolled-iteration type reconstruction methods, which can be considered state-of-the-art as per Leuschner et al. [LSG+21], is the Primal-Dual Network [AÖ18].

It attempts to imitate the non-linear primal dual hybrid gradient (PDHG) algorithm [CP11], which is able to solve inverse problems of potentially non-smooth objective functionals iteratively by replacing the exact gradient calculation needed for the optimization with proximal operators that are differentiable and directly calculate one step of the optimization. Additionally, the optimization is carried out in both the primal space, i.e. which the variable to be optimized is an element of, as well as in the dual space, i.e. the range of the operator which is to be inverted, alternatingly.

Instead of hand-crafting the proximal operators, they are defined as trainable layers

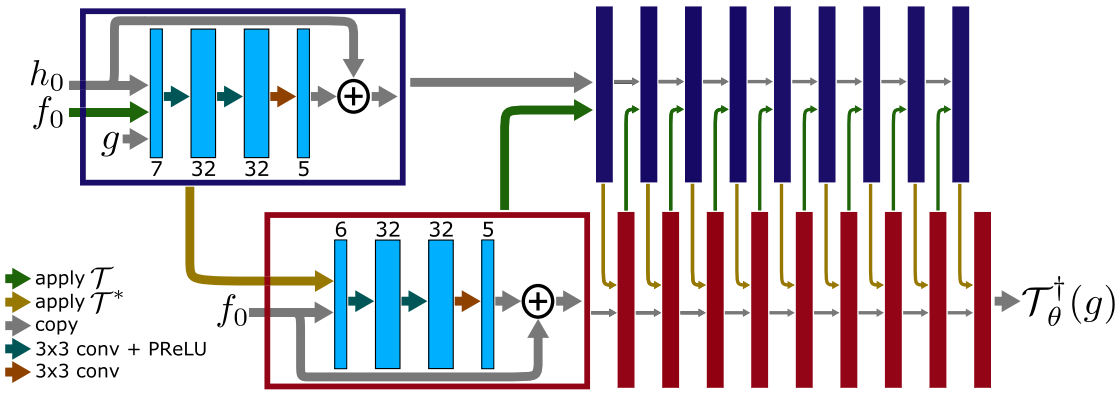


Figure 4.3: Primal-Dual Network: an unrolled iterative reconstruction method with CNNs [AÖ18] © 2018 IEEE.

of a neural network which can be optimized in a supervised manner depending on the data that the algorithm should be applied on afterwards.

Moreover, the presented *Learned Primal-Dual* algorithm generalizes the learned PDHG algorithm by modifying some steps (keeping track of the data in between the iterations, making the update in the dual space trainable instead of a constant weighted summation, letting the network learn the over-relaxation for the primal updates instead of a constant factor and allowing to learn different proximal operators in each iteration).

The Primal-Dual Network which they proposed (see Fig. 4.3) consists of residual convolutional blocks in both primal and dual space. The unrolled iteration count was set to 10 and the primal variables at the beginning were zero-initialized, since other initializations, e.g. FBP reconstructions, did not give better final results on their data.

Though outperforming other reconstruction algorithms (both algebraic and trainable ones), the architecture does not scale well: each additional unrolled iteration increases not only the processing time (especially when the operators for switching between the primal and dual space are not highly optimized, e.g. Radon transform and FBP) but also the memory consumption (especially when applied to problems in a higher dimension, e.g. three-dimensional CT volume reconstruction), because the intermediate variables need to be stored for each iteration.

4.3 Temporal and Model Prior Knowledge

As described in Sec. 3.6, iterative reconstruction methods are the type of algorithms that can incorporate this type of prior knowledge most easily. Therefore, most of this prior knowledge, expressed as (additional) terms in a cost function for optimization, are found in publications presenting algebraic or statistical reconstruction algorithms. Moreover, since the training of deep learning models is a type of (stochastic) gradient descent, and therefore iterative, several loss functions for neural networks include these additional terms, as well.

One of the most fundamental Model Prior Knowledge about reconstruction tasks is the representation of the data. As described in Sec. 3.5, projection data from a real CT scanner is a finite set of measurements (due to the discrete nature of the detector pixels). However, this does not specify how the reconstructed data is represented, i.e. how the actual continuous function of attenuation coefficients is discretized. The most obvious, and probably mostly used, representation is pixels/voxels. However, assuming the discretization of the continuous object function to be a weighted sum of basis functions, other representations become possible, too. Next to pixels/voxels, Kaiser-Bessel functions as basis functions are conceivable, which represent blobs, and are chosen for their favorable mathematical properties (rotational symmetry and relatively low computational cost compared to basis functions other than pixels/voxels) [NWV$^+$15][ZNG08].

Especially traditional (non-deep learning) reconstruction algorithms and deep learning networks incorporating projection/reconstruction layers need to define how the detector data should be interpreted: as an integral (from the source to the detector pixel) over a line [Sid85], a strip (effectively a rectangle or cube), or trapezoids/truncated pyramids [LFB10] (sorted by increasing degree of reality and decreasing computation speed). While trapezoid integrals can be considered state-of-the-art for CT reconstruction in general, projection/reconstruction layers in deep learning networks still mostly rely on line integrals for their computation speed.

Another choice about the physical representation, for statistical reconstruction in particular, is how to model the type of the probability distribution of the detector pixel values and the corresponding noise. In fact, the noise is a combination of Poisson and Gaussian distributions [TBS$^+$06] but due to mathematical properties, this is often reduced to only Poisson [HL89] or only Gaussian noise [BS93]. Very often, though, noise is not considered at all if the publication's goal is to show how a certain aspect of an algorithm behaves under simplified/idealized conditions. The presented methods in Ch. 5 will not consider photon noise, either, as the main focus of the thesis is reduction of streaking artifacts caused by sparse views.

Although the previously mentioned choices are important for modeling and representing the physical data, the definition of Model Prior Knowledge is more general in mathematical optimization and the terminology depends on the field of application, e.g. optimizing an *objective function* in mathematics, an *energy function* in physics or a *loss* or *cost function* in machine learning. Since this thesis focuses on CNN based algorithms, being a part of machine learning, the term *loss function* will be used. The loss function is generally defined by two sub-terms [Bis06, Sec. 3.1.4]:

$$GeneralLoss(F, G) = \Phi(F, G) + \lambda\Psi(F), \quad \lambda \in \mathbb{R}_{>0}$$

where Φ is a data mismatch function (a.k.a. data term, Sec. 3.8) and Ψ is a regularization function (a.k.a. model term) attempting to reduce the nullspace and therefore generating both more unique and credible solutions. Compared to the data term, the model term only depends on the output of the reconstruction algorithm (and not additionally on the ground truth), which means it can only measure inherent errors in the reconstructions.

Common choices for model terms include Huber [ZZH$^+$13], total variation [SP08],

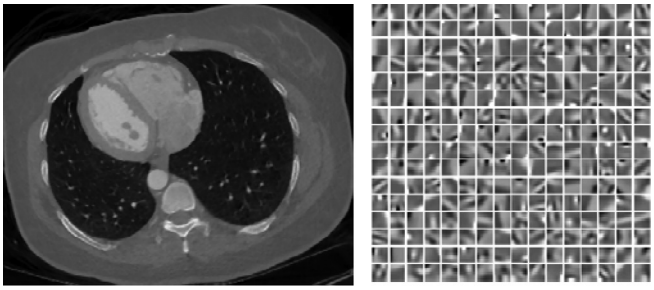


Figure 4.4: Global dictionary learning [XYM+12] for low dose CT reconstruction. (a) Slice used to extract the patches. (b) Learned dictionary consisting of 256 patches © 2012 IEEE.

and pseudo L0 norm [HXL+11] regularization. Moreover, model terms can also express Temporal Prior Knowledge, i.e. knowledge about the scanned subject from, e.g., high quality planning scans, projections that were acquired earlier during surgery or information about locations of interventional instruments which are tracked over time. Examples include predicted Markov random field coefficients [ZHL+16], global dictionary learning [XYM+12] (see Fig. 4.4) or non-local means regularization [ZHM+14]. For a more complete review of regularization functions, the reader is referred to Zhang et al. [ZWZ+18].

It is not possible to determine a single state-of-the-art method in the scope of Temporal and Model Prior Knowledge because the presented model terms are specific to certain tasks to be solved. As an example, a highly weighted total variation term effectively reduces or even eliminates artifacts like noise or streaks, such that a subsequent segmentation is greatly simplified. However, using the very same weighted term for just an enhanced reconstruction using less X-radiation likely simplifies textures of tissues (i.e. reduces local contrasts) excessively such that a distinction of tissues with similar average attenuation coefficients but different contrasts might become very challenging.

4.4 Discussion

Summarizing the previous sections, it can be concluded that there is not one particular method that can be considered state of the art in terms of CT reconstruction. While many of the rather general algorithms, e.g. FBP or FBPConvNet, work reasonably well for all kinds of reconstructions, the more specialized ones, e.g. Primal-Dual Network, may solve certain reconstruction problems best but have not been evaluated on other less specialized problems or reconstruction tasks of different domains.

Moreover, one main problem of the methods and architectures explained in the section about Machine Learning Prior Knowledge is their scalability. On one hand, this is often less of a problem in case of sparse view CT when the methods process the projection images. However, the projections as well as reconstruction images/volumes were in

all cases described to be of low resolution, i.e. lower than the actual pixel count of a real detector or lower than the actual pixel/voxel count of a reconstruction generated from a real CT system. The limiting factors are two-fold: not only does the memory consumption of high-resolution reconstruction still quickly exceed current graphics cards' RAM limitations, but the processing time also is usually proportional to the data size, which might not be a big problem during inference but easily increases the training time to impractical amounts. Therefore, the applicability of the machine-learning-based reconstructions is limited to use cases that do not rely on highly accurate images of especially very small structures.

Another very important brick missing in the wall of CT reconstruction methods aided by machine learning is algorithms for (interventional) cone beam CT reconstruction. Despite a large amount of CT reconstruction networks, most of them were designed for parallel beam CT, which is nearly unused for medical imaging nowadays, or for conventional CT, including fan beam geometries or rather thin cone beam projections from multi-row detectors. Technically, the methods developed for these geometries could also be applied to CBCT acquisitions from flat panel detectors, but would result in discarding many detector rows that are not close to the central row and therefore invalidate the assumption of projections perpendicular to the transverse axis, which many reconstruction methods make. True CBCT methods should make use of all detector rows, including those with a high cone angle. This, however, usually results in a much higher memory consumption unsuitable for deep learning. For this reason, additional processing of the cone beam projections or modifications to the reconstruction algorithms, i.e. Algebraic Prior Knowledge, are necessary.

Lastly, there is no evaluation in the literature about which types of prior knowledge aid the reconstruction to what extent. Similarly, it has not yet been assessed if combinations of different types of prior knowledge may reinforce each other or if one type is already sufficient for certain use cases.

Chapter 5

Methods

In this chapter, the methods that were developed and evaluated are described and explained in detail. Each section will start with an introducing paragraph which supplies further information about how the method fits in this thesis and which publication it is based on. Moreover, the main types of prior knowledge used for the respective methods will be stated in this introductory paragraph as well. This way, the reader is immediately aware of where to place the methods in the scope of this thesis.

Before describing the actual methods, though, the first section of this chapter will introduce the data sets that were used to train and evaluate the methods. This is to get a first visual impression of how the data looks like and to provide some further information about the imaging parameters and statistics about the scanned population. In the description of the methods, the data sets merely need to be referenced and possible selections of subsets need to be justified.

5.1 Data Sets

For neural networks, the data sets that are used for training are an essential part to achieving good quality as well as good generalization. The task of CT image reconstruction is not limited to certain anatomical sites of the human body nor to medical imaging in general. However, since this thesis focuses on medical CT reconstruction, the data sets will be selected to cover certain areas of the human body that are often scanned in clinical routine or therapies, or that have certain properties (like untruncated projections) such that some types of artifacts or other problems do not distort the capabilities of the presented methods and to limit the space of solutions. However, this does not mean that the data sets were cherry-picked and the algorithms only work on a very limited subset of the available data, but it should be kept in mind that modifications to the methods might be necessary when being applied to new data.

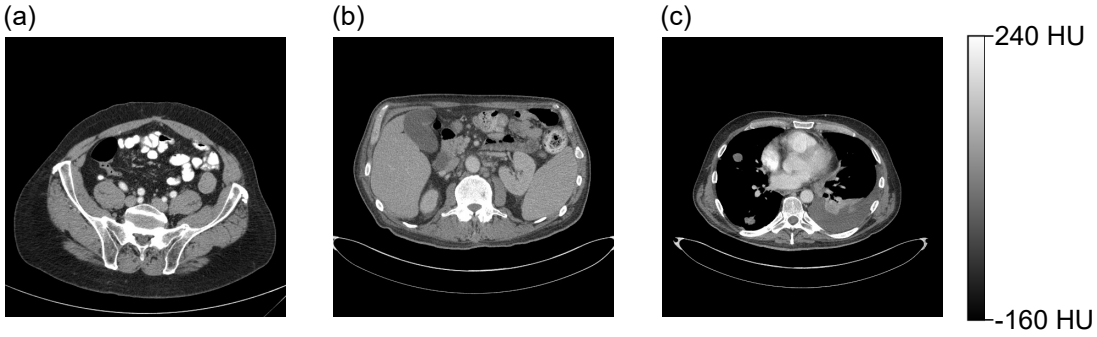


Figure 5.1: Three exemplary axial slices of the CT Lymph Nodes data set from three different subjects. (a) Pelvis. (b) Midabdomen. (c) Mediastinum.

5.1.1 CT Lymph Nodes

The CT Lymph Nodes collection [RLS+15] is a publicly available data set of The Cancer Imaging Archive [CVS+13] which comprises 176 CT scans of different patients of the mediastinal and abdominal region (see Fig. 5.1). It has been originally designed for developing and assessing automated detection algorithms of lymph nodes, which is a challenging task due to the high variance in shape and size and low contrast of surrounding tissues. The provided data is the reconstructed CT volumes in Hounsfield units. The axial in-plane resolution of the scans is $512\,\mathrm{px} \times 512\,\mathrm{px}$ with varying pixel spacings (between $0.664\,\mathrm{mm} \times 0.664\,\mathrm{mm}$ and $0.977\,\mathrm{mm} \times 0.977\,\mathrm{mm}$, the median being $0.803\,\mathrm{mm} \times 0.803\,\mathrm{mm}$) and varying numbers of slices (between 485 and 746 with a median of 674 slices). The spacing between slices is usually 1 mm and occasionally 1.25 mm. Unfortunately, there is no information about the X-ray tube, detector or gantry parameters.

Since the scans were acquired using a conventional CT system and the main focus of this thesis is on interventional CBCT, it is not problematic that the projections are not included in this data set.

The data set also includes labels, i.e. positions of the centroids as well as size measurements and segmentations, for 388 mediastinal and 595 abdominal lymph nodes. However, these are not going to be used in the scope of this thesis.

This data set was selected because it comprises a great variety of subjects wrt. age (between 18 and 73 years) while being stratified wrt. male/female. Moreover, the volumes comprise a wide range of pixel spacings and spacings between slices. This makes them ideal as training data for CNNs in order to generalize well and be applicable to new data.

5.1.2 NeuWave Medical Needle

This data set comprises several scans of a NeuWave Medical ablation needle (see Fig. 5.2).

The needle was inserted into an abdominal CT phantom at different positions and various angles. These combinations of needle and phantom were then subsequently

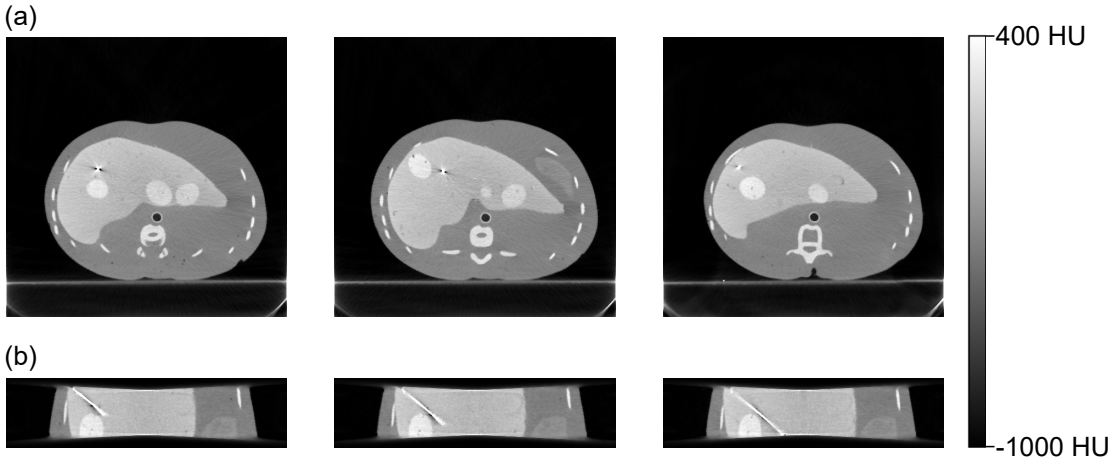


Figure 5.2: Three exemplary scans of the NeuWave Medical Needle data set. (a) Axial slices. (b) Coronal slices, included here to visualize the extents of the needle.

scanned in the in-house KIDS-CT at *STIMULATE*, a conventional CT scanner with a cylindrical detector and a tubular gantry, with an axial acquisition protocol. The tube's peak voltage was set to 100 kV with an exposure of 42 mAs for each X-ray projection.

Due to the axial acquisition protocol and the rather small (in terms of rows) detector, the field of view was rather narrow with about 60 mm along the transverse axis. Therefore, the reconstructions were created with an axial in-plane resolution of 512 px × 512 px with a pixel spacing of 0.75 mm × 0.75 mm for a number of 128 slices with a slice thickness of 0.5 mm each.

This data set was only used to extract interventional needles which could then be inserted virtually into the scans of the other described data sets. To this end, the reconstructions were segmented using a simple thresholding of 2500 HU which perfectly separated the needle from the phantom values (disregarding small artifacts very close to the needle). These volumes containing only the needles were then reprojected to simulate interventional CBCT projections that could be combined with simulated CBCT projections of the other data sets.

The lack of large data sets of interventional CBCT scans including needles makes this simulation a necessary step.

It should be noted that the extraction of the needle always needs the indirection of a reconstruction, even if scanned in a CBCT system, unless the needle is held in position without any aid (like the phantom, in this case). This is because the projections are ray sums and therefore, extracting the needle directly on the projections does not suppress the attenuation values of the surrounding materials on the corresponding rays.

This data set was selected because it is one of the few available data sets depicting interventional instruments. As discussed earlier, scans acquired during medical interventions are rarely saved for later reference due to their low quality compared to conventional CT scans and their therefore rather limited number of use cases after surgery.

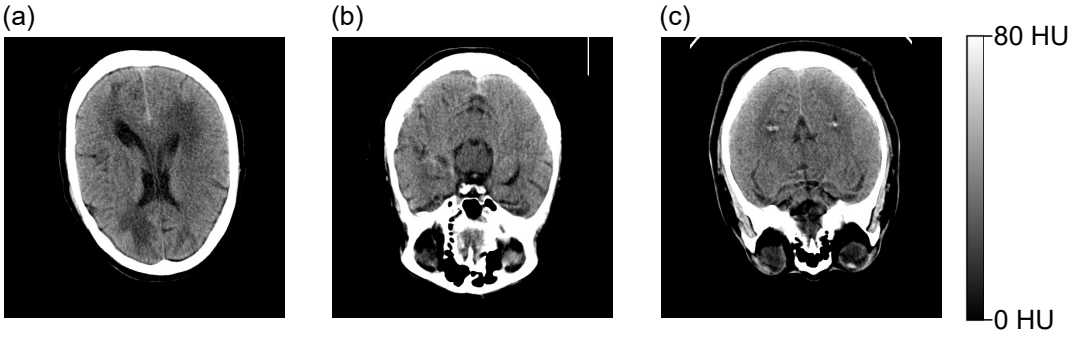


Figure 5.3: Three exemplary axial head scan slices of the May Clinic data set from three different subjects.

5.1.3 Mayo Clinic Data Set

The Mayo Clinic Data Set [MCH+20] is a publicly available data set, originally designed for the Low Dose CT Grand Challenge, the objectives of which being the qualitative assessment of the diagnostic performance of denoising and iterative reconstruction algorithms. It comprises 299 scans of the head, chest and abdomen, half of them from a Siemens SOMATOM Definition Flash scanner, the rest from a GE Lightspeed VCT CT scanner (see Fig. 5.3).

It does not only include the reconstructed volumes of anonymized clinical routine dose acquisitions as well as simulated low dose reconstructions but also the raw X-ray projection data, which is especially helpful for developing reconstruction algorithms that do not solely rely on post-processing the reconstructions but directly incorporating the projections without the need to simulate this raw data. This way, new algorithms can be developed for and tested on actual real data.

Additionally, the data set also contains an Excel sheet with clinical data identifying all pathology, as well as annotations of lesions.

Like the CT Lymph Nodes collection, the Maya Clinic Data Set has been available on The Cancer Imaging Archive since 2020, i.e. four years after the challenge took place, in a restricted area and needs a free registration before being able to access. However, the data is now also available on the challenge website without needing to register.

Only fifty of the available 99 head scans were used in this thesis for memory reasons.

Again, the projections provided in this data set could not be used for the purpose of this thesis since they were acquired from conventional CT systems and not from interventional CBCT.

This data set was used because of the favorable extents of the scanned anatomic site: the head. Since the flat panel detectors of interventional C-arm scanners are approximately the size of a human head, the projections are merely truncated caudally whereas the projections of the sites of the previously described data sets are truncated in both cranial and caudal directions, and possibly laterally as well. This lack of truncation of the head scans results in reconstructions which bear fewer artifacts such that the

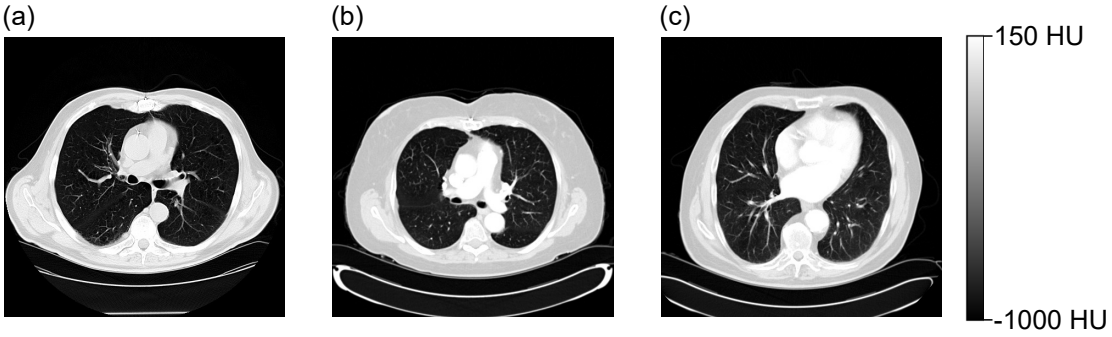


Figure 5.4: Three exemplary axial head scan slices of the LungCT-Diagnosis data set from three different subjects.

trained networks can focus on optimizing the artifacts caused by sparse views and highly absorbing materials like ablation needles.

However, the application of networks trained for sparse view interventional head scans for, e.g., tumor ablation is rather limited because brain tumors are usually more fatal than a higher dose of X-radiation such that, in this case, surgeons would rather increase the X-ray exposure to be able to remove the tumor more accurately with higher quality reconstructions in the first place.

Nevertheless, one can expect methods that work well on this data set to work well for different anatomical sites, too: the dynamic range of attenuation coefficients in the cranial region is among the highest in the entire human body. It comprises highly absorbing tissues like the skull as well as soft tissues like the brain, which makes the reconstruction of cranial acquisitions a challenging task.

5.1.4 LungCT-Diagnosis

The LungCT-Diagnosis data set [GBS+15] is another publicly available data set on The Cancer Imaging Archive [CVS+13]. It comprises 61 retrospectively, routinely acquired diagnostic helical CT scans with enhanced contrast (see Fig. 5.4). Its original purpose was to develop methods for quantitatively describing lung adenocarcinomas from CT acquisitions.

Out of the 61 scans, 54, 5, and 2 were acquired with a Siemens, GE Medical Systems and Toshiba scanner, respectively. The axial in-plane resolution is 512 px × 512 px with varying pixel spacings (between 0.586 mm × 0.586 mm and 0.953 mm × 0.953 mm with a median of 0.725 mm × 0.725 mm) and varying numbers of slices (between 24 and 150, the median being 70 slices). The slice thickness ranges from 3 mm to 6 mm. The peak voltage of the X-ray tube is 120 kV in 57 cases, 130 kV in 1 case, and 140 kV in 3 cases. The average exposure per slice ranges from 66 mAs to 313 mAs with a median of 110 mAs. All image data is represented as Hounsfield units. Further information about the scanning geometry can be found in the headers of the image files.

Along with the image data, there is an additional document summarizing the statis-

tics of the scanned patients and the imaging parameters, as well as clinical parameters for each patient of the cohort.

There is no raw/projection data in this data set. However, like for the other publicly available data sets, the scans were acquired using conventional CT scanners, so the lack of projections is not problematic for the methods that are designed for CBCT.

This data set was selected because of the large number of use cases of high quality despite low dose reconstructions of this anatomical site. Radiofrequency ablations of lung tumors are a minimally invasive alternative to open surgery [JFS+13] usually guided by CBCTs imaging, and chest CT scans in general have become more frequent during the past years [WMM+20; HHL+16], the coronavirus disease 2019 having a significant impact on hospitals' imaging department's routine activity [BBF20]. Developing methods that improve the quality of the reconstructions in this anatomical site is therefore much appreciated.

5.2 Deep Learning Prior Knowledge: Primal-Dual UNet for Sinogram Upsampling

This section is in large parts based on the publication "Sinogram upsampling using Primal-Dual UNet for undersampled CT and radial MRI reconstruction" [ECR+21] in collaboration with Soumick Chatterjee. Here, it will be described for sparse view fan beam CT data but it has also been shown to work well for undersampled radial MRI acquisitions and is in many cases even better than using the traditional reconstruction algorithms of MRI. These additional results can be found in the appendix.

The prior knowledge used in this method is primarily categorized as Deep Learning Prior Knowledge *and secondarily as* Analytical Prior Knowledge*.*

5.2.1 Architecture: Primal-Dual UNet

The network architecture proposed in this paper is based on the Primal-Dual Network [AÖ18], which achieves superior reconstruction quality compared to other deep learning based reconstruction algorithms [LSG+21], and is combined with UNet [RFB15b], which is well-known and often used in medical image processing.

The Primal-Dual Network can be interpreted as an unrolled iterative reconstruction algorithm that optimizes in both the sinogram and image space using blocks of fully convolutional layers. In each iteration, the processed sinograms are reconstructed using FBP, combined with the processed images, and reprojected to be combined with the previous sinograms. This does not only improve the quality of the final reconstructions, but also ensures data consistency with the original sinogram. As with many iterative algorithms, the quality of the reconstructions of the Primal-Dual Network depends on the number of iterations, i.e. for a fixed number of parameters in an iteration, an optimal (usually minimal) number of iterations has to be found for the network to converge. If the number of parameters in the convolutional blocks is low, more iterations are needed for convergence. This, however, increases the processing time, with the reconstruction

and projection operators being the bottleneck. On the other hand, if the number of parameters in the convolutional blocks increases, much fewer iterations are needed for convergence. The processing time is not much different though, since the convolutions are the bottleneck in this case.

The idea of the network architecture proposed in this paper tackles this trade-off by

1. keeping the number of iterations low, i.e. the reconstruction and projection operators are not a bottleneck, and

2. replacing the convolutional blocks in the image space with a UNet, to get a high number of parameters while keeping the processing time low.

This architecture will be referred to as *Primal-Dual UNet*, or short *PD-UNet*.

5.2.2 Baselines

The performance of the proposed model was compared against three deep learning based methods. First, a UNet [RFB15b] was applied on reconstructed undersampled CT images [JMF$^+$17] and MRIs [HKL$^+$18], referred to hereafter as *Reconstruction UNet*. The CT images were reconstructed using FBP, and the undersampled MRIs were reconstructed using adjoint NUFFT [Lin18]. The undersampled images were supplied as input to the Reconstruction UNet model, and the outputs of the model were compared against the ground-truth fully-sampled images to calculate the loss while training.

The second baseline was the *Sinogram UNet* [LLK$^+$19]. For this method, the sparsely sampled sinograms were upsampled using bilinear interpolation before supplying them to the UNet model as input, and the loss was computed by comparing the outputs of the model against the corresponding fully-sampled sinograms.

The final deep learning baseline was the *Primal-Dual Network* [AÖ18], where the sparsely sampled sinograms and zero-initialized reconstructions were supplied as input to the model, and the outputs from the model were compared against the ground-truth fully-sampled images to calculate the loss.

Lastly, the performance of the models was compared with FBP reconstructions after upsampling the sparse-sampled sinograms using bilinear interpolation.

5.2.3 Data Normalization

Data in the image space was normalized by dividing each slice by the 99^{th} percentile of the intensity values present in the total training and validation sets. The sinograms were normalized using the Z-score normalization method – by applying the following equation on each sinogram:

$$S_N = \frac{S - \mu_s}{\sigma_s} \tag{5.1}$$

where S_N is the normalized sinogram, S is the original sinogram, and μ_s and σ_s are the mean and standard deviation of the values present in the sinogram.

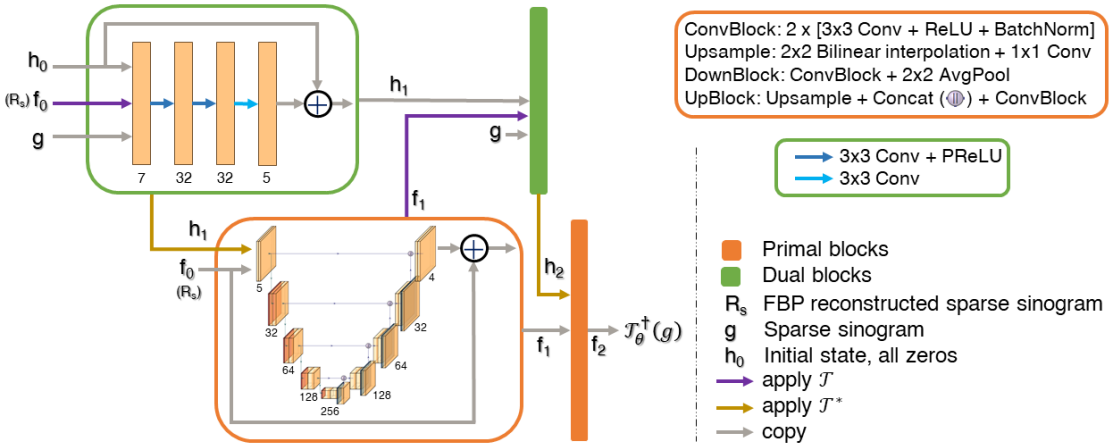


Figure 5.5: Primal-Dual UNet (PD-UNet): Architecture of the proposed network. The orange and the green boxes signify the primal and dual iterates, respectively. The proposed primal block uses a UNet model instead of a fully-convolutional network. In contrast, the dual block is the same fully-convolutional network as the original primal-dual network. It is to be noted that the second orange and green blocks have the same architecture as the first one. The initial dual block receives the image reconstructed by applying FBP on the sparse sinogram as f_0 (unlike the original primal-dual network, which receives all zeros), all-zeros as the initial-state h_0, and the sparse sinogram g. The output of this block, along with the same f_0 is given as input to the first primal block.

For the Reconstruction UNet and Sinogram UNet, the input was normalized using the image space normalization and sinogram normalization methods, respectively. The output given by those models were "de-normalized" to obtain the final output – for image space, multiplying with the 99^{th} percentile of the intensity values; for sinogram space, by applying the following equation:

$$S = (S_N \times \sigma_s) + \mu_s \tag{5.2}$$

For the Primal-Dual Network and the proposed PD-UNet, several normalization and de-normalization steps were performed. Initially, μ_s and σ_s were calculated from the sparsely sampled input sinogram, and all the sinogram (de-)normalizations were performed using these values. Each time before a sinogram or an image was given as input to any block, they were normalized using sinogram and image normalization methods as discussed earlier, and after receiving output from that block, they were de-normalized. This was performed to preserve the relationship between image and sinogram values while using two different types of normalization techniques for two different data spaces. Before providing the final output of these models, the values were also de-normalized using the image space technique.

5.2.4 Implementation and Training

The models (the proposed model and the baseline models) were trained for 151 epochs with an effective batch size of 32, the best epoch was chosen based on the validation loss and was used for inference on the test set. The memory requirements of the proposed model and the different baseline models are not the same, making it impossible to have the same batch size for the different models. To achieve a constant effective batch size for all the models, instead of the conventional "forward-pass then backward-pass" technique for each training step, multiple forward passes were performed, the gradients were summed up, and finally, an accumulated backward-pass was performed. The number of forward passes to be performed before an accumulated backward pass was calculated as: (32 ÷ actual batch size of the model). To train the models, the loss was calculated using mean absolute error (L1-Loss), and it was minimized using Adam optimizer (learning rate $= 1 \times 10^{-3}$, $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1 \times 10^{-8}$). The models were implemented using PyTorch [PGM+19], with the help of PyTorch Lightning [Fal+19]; and were trained with mixed precision [MNA+17] using Nvidia RTX 2080Ti and Nvidia RTX A6000 GPUs. The code of this project is available on GitHub[1].

5.2.5 Data Set

In this study, the data of 28 subjects of the CT Lymph Nodes [RLS+15] collection from The Cancer Imaging Archive [CVS+13] was used, consisting of reconstructed volumes of the abdomen that serve as ground truth. 16, 4 and 8 subjects were used for training, validation and test set, respectively. Since the number of axial slices and the voxel sizes

[1] `https://github.com/phernst/pd-unet`

differed per subject, the central 200 slices of each subject were extracted and interpolated using sinc interpolation to have an in-plane matrix size of 256x256.

The sinograms were simulated using torch_radon [Ron20] as the data sets only contain reconstructions. Each fan-beam projection consists of 511 detector pixels with a spacing of 1 px, a source-to-isocenter distance of 400 px and a detector-to-isocenter distance of 150 px to cover the full axial slice. The sinograms contain 360 equiangular projections with an angular distance of 1° between consecutive projections. The reason for choosing fan-beam over parallel-beam projections here is its applicability to real-world data. However, results for parallel-beam projections can be found in Appendix A.4.

5.2.6 Undersampling

To simulate the undersampled data sets, the sinograms were made sparse by retaining only every n^{th} projection, where n denotes the level of sparsity. For CT, three different sparsity levels were experimented with: 4, 8 and 16 (referred herewith as Sparse 4, Sparse 8 and Sparse 16, respectively).

5.2.7 Evaluation criteria

The performance of the models was evaluated and compared quantitatively with the help of RMSE and SSIM [WBS+04]. Moreover, the statistical significance of the improvements observed was evaluated by the Mann–Whitney U test [MW47]. Finally, they were also compared qualitatively for selective slices with the help of difference images and SSIM maps. Slices that resulted in SSIM values identical to the median value in up to three decimal points for Primal-Dual UNet (the proposed method) and Primal-Dual Network (main baseline) were chosen for qualitative portrayal for each experiment - to be able to choose slices which are representative of the results for each of the experiments.

5.2.8 Results

The performance of the proposed Primal-Dual UNet was compared (see Sec. 5.2.2) both quantitatively and qualitatively against three other deep learning models: Reconstruction UNet [JMF+17], Sinogram UNet [LLK+19], and Learned Primal-Dual Network [AÖ18], and also against reconstructions with the standard FBP applied on the sinograms up-sampled using bilinear interpolation, referred to here as *Sinogram Bilinear*. The fan-beam geometry being more in more real-world use than the parallel-beam geometry, the focus of this research was on the fan-beam geometry - hence these results are shown in this section. Additional experiments were also performed with the parallel-beam geometry, and the results have been reported in Appendix A.4. Experiments were performed for three different levels of sparsity: 4, 8 and 16 (referred to as: Sparse 4, 8 and 16, respectively).

Quantitative evaluations were performed using RMSE, calculated in the Hounsfield scale, and SSIM, calculated on the normalized intensity values, as shown in Tab. 5.1.

Table 5.1: Resultant metrics for CT fan-beam geometry (mean±std)

Method	SSIM		
	Sparse 4	Sparse 8	Sparse 16
Bilinear Sinogram	0.928±0.011	0.824±0.021	0.716±0.033
Sinogram UNet	0.977±0.005	0.948±0.016	0.874±0.032
Reconstruction UNet	0.983±0.003	0.953±0.012	0.903±0.026
Primal-Dual Network	0.983±0.003	**0.973±0.005**	0.919±0.016
Primal-Dual UNet	**0.985±0.002**	0.966±0.008	**0.932±0.021**

Method	RMSE (Hounsfield units, HU)		
	Sparse 4	Sparse 8	Sparse 16
Bilinear Sinogram	33.135±4.557	59.588±7.645	90.148±12.033
Sinogram UNet	14.482±2.324	26.811±13.810	47.574±11.621
Reconstruction UNet	**11.860±1.891**	25.575±5.384	47.689±12.170
Primal-Dual Network	21.693±3.216	23.868±3.806	35.386±6.212
Primal-Dual UNet	15.835±2.143	**22.343±4.367**	**34.383±8.788**

The range of the resultant SSIM values is portrayed with the help of box plots in Fig. 5.6 for the three different levels of sparsity.

In terms of SSIM, the proposed method achieved improvements over all the baseline methods with statistical significance for Sparse 4 and 16 - including improvements of 0.2% and 1.39% respectively over the baseline Primal-Dual Network. However, Primal-Dual Network scored 0.72% better average SSIM than the proposed method with a statistical significance for Sparse 8. On the other hand, the proposed method scored better RMSEs (27%, 6.39%, 2.83% for Sparse 4, 8, 16) than the baseline Primal-Dual Network with statistical significance for all three levels of sparsities. For Sparse 8 and 16, the proposed method achieved better RMSEs than all the baselines. However, for Sparse 4, both Sinogram UNet and Reconstruction UNet achieved better RMSEs than the baseline Primal-Dual Network, as well as the proposed method.

Fig. 5.7 shows qualitative comparisons of the reconstructions for Sparse 8 and 16, respectively. Comparisons are performed with the help of difference images (in the Hounsfield scale) and SSIM maps (calculated on the normalized intensity values). By looking at the qualitative results, it can be said that they do corroborate with the quantitative results.

Simulated needle insertion

As a further test on the practical use of the proposed method, the insertion of an interventional needle into the abdominal scans is simulated and evaluated visually and

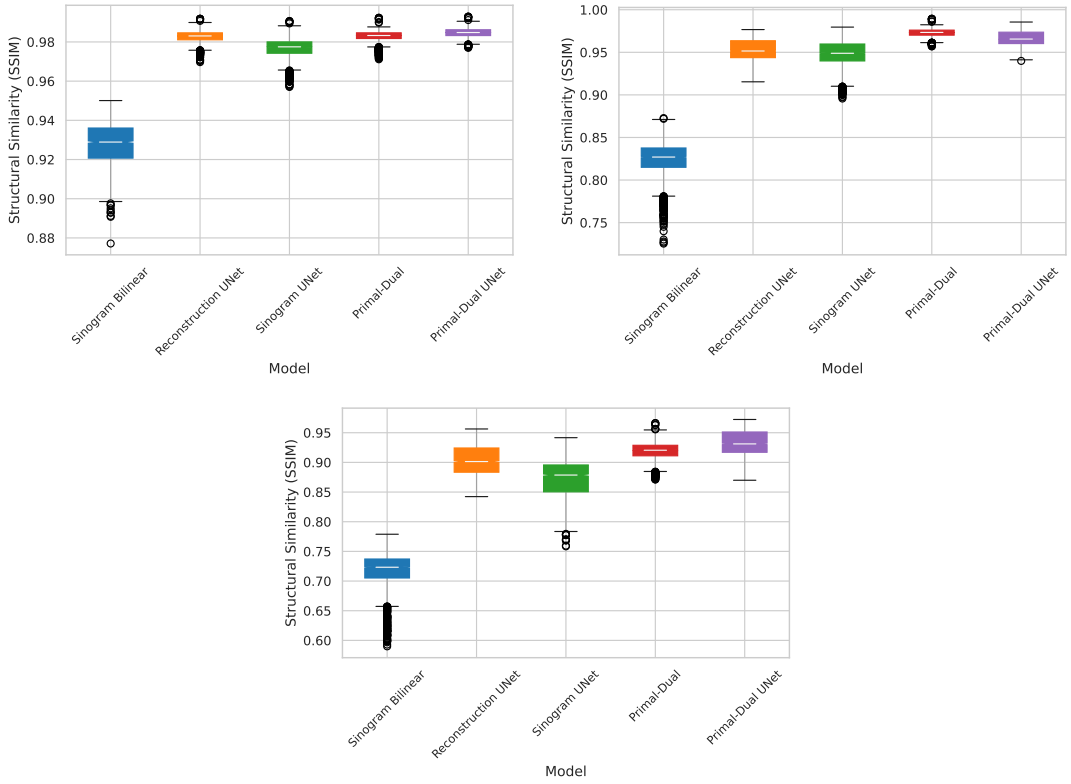


Figure 5.6: Box-plots of the resultant SSIM values for CT (fan-beam geometry) Sparse 4 (top left), Sparse 8 (top right) and Sparse 16 (bottom)

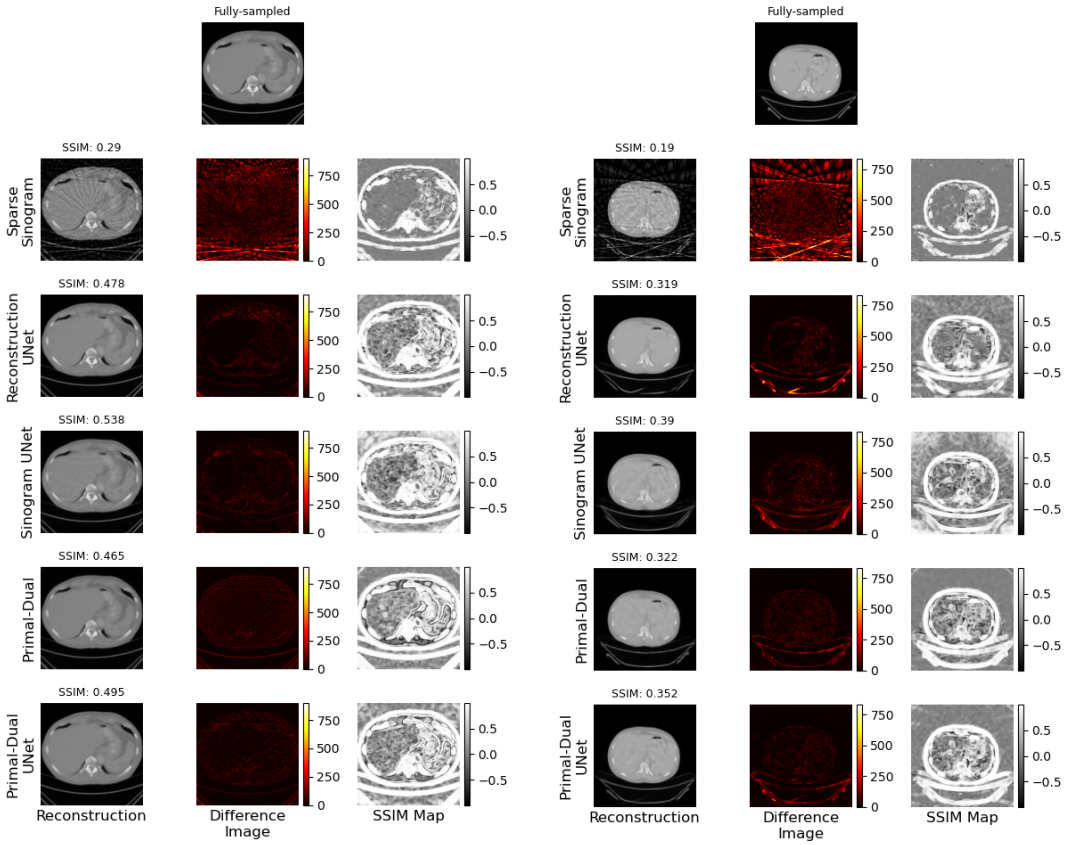


Figure 5.7: Qualitative comparisons of the reconstructions for fan-beam geometry (left: Sparse 8, right: Sparse 16)

Table 5.2: Resultant metrics for CT fan-beam geometry with inserted needle for Sparse 16 (mean±std)

Method	SSIM	RMSE (HU)
Sparse Sinogram	0.817±0.015	355±8
Bilinear Sinogram	0.871±0.006	523±2
Sinogram UNet	0.915±0.008	430±9
Reconstruction UNet	0.919±0.011	403±3
Primal-Dual Network	0.913±0.005	458±6
Primal-Dual UNet	**0.940±0.004**	**367±8**

quantitatively. For this purpose, a NeuWave Medical[2] ablation needle was inserted into an abdominal phantom and was scanned with a KIDS-CT scanner. The needle was segmented out of the resulting volume by a simple thresholding. The needle was combined with the available test volumes by summing the attenuation coefficients, which represents a good estimation of actual needle insertion since the values of human tissues and needle materials are significantly different. However, this simulation is missing some artifacts caused by, e.g., photon starvation. These combined volumes served as the ground truth for this experiment. Sparse fan-beam sinograms and reconstructions were again simulated using pytorch_radon, and the same pre-processing was performed as before. Qualitative and quantitative results of an exemplary slice for Sparse 16 are shown in Fig. 5.8. Despite not being trained on data sets with needles, the networks seem to be capable of reconstructing these highly absorbing materials instead of assuming them to be artifacts to be removed or replaced by soft tissue attenuation coefficients. Similar to the results obtained in the previous experiments without the needle, all networks improve the FBP reconstruction of the sparse sinogram by at least 25 percentage points SSIM. Sinogram UNet still performs worst, followed by Reconstruction UNet, Primal-Dual Network, and the best performing Primal-Dual UNet with 0.979, 0.981, 0.985, and 0.987, respectively.

It is of special interest to evaluate the reconstruction in the region around the needle. For this reason, a 32x32 patch around the needle was extracted from every prediction of all test volumes, and the errors were calculated on this region of interest only. These results can be found in Tab. 5.2.

The trend continues as described previously: a simple upsampling of the sinogram results in a small increase of 5 percentage points SSIM wrt. the FBP of the sparse sinogram. Interestingly, Reconstruction UNet performs slightly better in terms of the average SSIM than Primal-Dual Network, though both still perform reasonably well with more than 0.91 SSIM. The proposed Primal-Dual UNet increases the SSIM of Reconstruction UNet by a large margin, to an average SSIM of 0.940 - an improvement of more than 2 percentage points. This shows that the presented model is not only capable of reconstructing images with higher quality compared to competing reconstruction

[2]Ablation needle: `http://www.neuwavemedical.com`

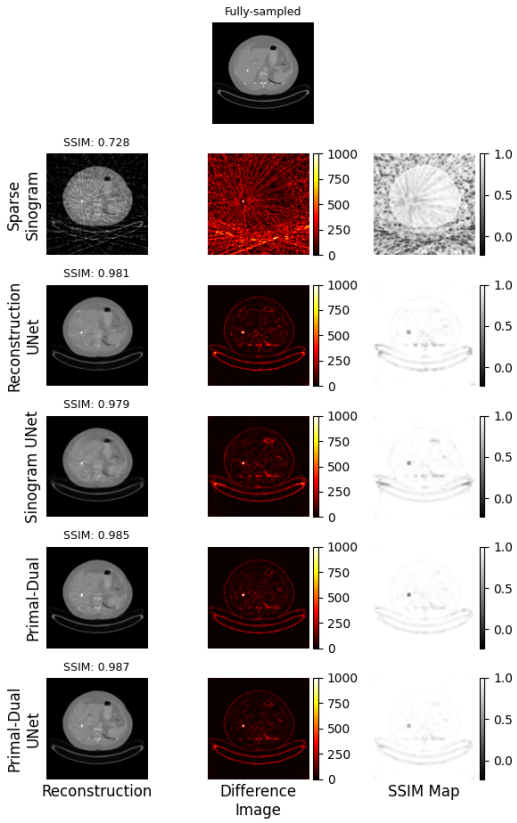


Figure 5.8: Qualitative comparisons of the reconstructions fan-beam geometry Sparse 16 with needle for an exemplary slice.

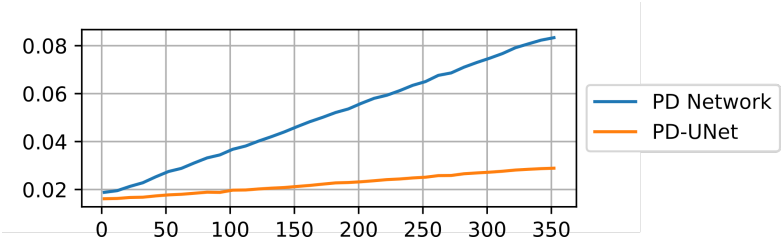


Figure 5.9: Comparison of inference times for Primal-Dual Network (PD Network) and Primal-Dual UNet (PD-UNet). Number of projections on horizontal axis. Average inference time for one batch (batch size: 4) in seconds on vertical axis.

networks trained on the same data set but is also likely to have a higher degree of generalization regarding different kinds of CT data sets and learns better how to reconstruct artifact-bearing CT data in general.

5.2.9 Comparison of Execution Speeds

Sparse sampling can reduce the speed of acquisition, which is an essential factor when it comes to MR imaging. However, the time required for reconstruction can be an additional overhead - increasing the total time for imaging. For this reason, the execution speed of the proposed Primal-Dual UNet was compared against the main baseline model - Primal-Dual Network. Fig. 5.9 shows the required amount of time to reconstruct one slice for these two methods and how much they change with a change in the number of projections. It can be observed that the proposed method is faster than the baseline Primal-Dual Network. Moreover, it can be observed that with the increase in the number of projections, the required reconstruction time increases for both models, but also the difference between the models increases constantly.

5.2.10 Discussion

The results revealed that all four deep learning based models performed better than applying FBP on the bilinearly upsampled sinograms. Sinogram UNet, which aims to refine those interpolated sinograms before applying FBP, performed the worst among the deep learning models. For the lowest level of sparsity (Sparse 4), Reconstruction UNet resulted in the same average SSIM as the main baseline of this paper - Primal-Dual Network - but resulted in a better average RMSE. However, the superiority of the Primal-Dual Network can be seen for the higher levels of sparsity. For Sparse 4 and 16, the proposed Primal-Dual UNet performed better than the baseline Primal-Dual Network. However, interestingly, the results of Sparse 8 are conflicting for average SSIM and RMSE. According to the average RMSE, Primal-Dual UNet outperformed the Primal-Dual Network - in accordance with the other sparsity levels, but resulted in a lower average SSIM than the baseline. However, as the Primal-Dual UNet performed better in five out of six scenarios, it can be concluded as the overall better-performing

model.

The needle insertion experiments showed that the networks not only learn to reconstruct images with tissues that they were trained on but also remove the artifacts from materials with significantly different (in this case: higher) attenuation coefficients without further training. As briefly described above, the volumes with the inserted needles were merely simulated and therefore lack other types of artifacts, e.g. caused by photon starvation (due to the high attenuation coefficients of the needles) or motion (due to the breathing of the subjects). Moreover, the patch size of 32x32 for the ROI evaluation around the needle was chosen empirically for this experiment to include the needle and some of the soft tissue which was affected most by the needle artifacts. Though this gives a good insight to how the different networks behave on unknown data, the quantitative values are susceptible to changes in the patch size and the shape of the depicted needle. Here, the shape was dot-like for the reconstructed axial planes, but it may as well be similar to a straight or bent line if inserted differently, which would invalidate the chosen patch size. In addition, conventional CT scanners nowadays usually use multirow detectors and helical acquisition trajectories, which was not taken care of in this study and might further increase the quality of the reconstructions.

5.3 Deep Learning Prior Knowledge: Primal-Dual UNet for Cone Beam CT Volume Reconstruction

This section in mainly based on the publication "Primal-Dual UNet for Sparse Cone Beam CT Volume Reconstruction" [ECR+22a] and serves as an extension to the method presented in the previous section.

The prior knowledge that is mainly used here is Deep Learning Prior Knowledge.

The main contributions of this work are:

1. modifying the Primal-Dual UNet (see Sec. 5.2) to process cone beam projections and

2. reconstructing entire volumes instead of axial slices.

5.3.1 Methods

The network architecture used in this work is a modified Primal-Dual UNet [ECR+22b]. The two-dimensional convolutions of the dual space blocks were replaced with their three-dimensional counterparts. The two-dimensional UNet in the primal space was replaced with a three-dimensional UNet by replacing convolutions, batch normalizations, average poolings and linear upsamplings with their three-dimensional counterparts. Instead of the parallel or fan beam projection layer, a cone beam geometry (detector: 310×240px, 1.232mm pixel size; SID=160mm; SDD=400mm) on a circular trajectory was used. The FBP reconstruction layer was replaced with its FDK counterpart.

Table 5.3: Mean and standard deviation over all axial test slices for *Sparse 16*.

Model/Method	SSIM [%]	PSNR [dB]	RMSE [HU]
FDK	43.54±8.27	17.92±2.64	388.57±108.15
FDKConvNet	67.37±8.99	24.72±1.93	177.30±60.94
Primal-Dual Network	69.87±7.66	25.19±2.08	169.22±64.35
Primal-Dual UNet	**78.76±7.50**	**27.93±2.33**	**128.89±57.78**

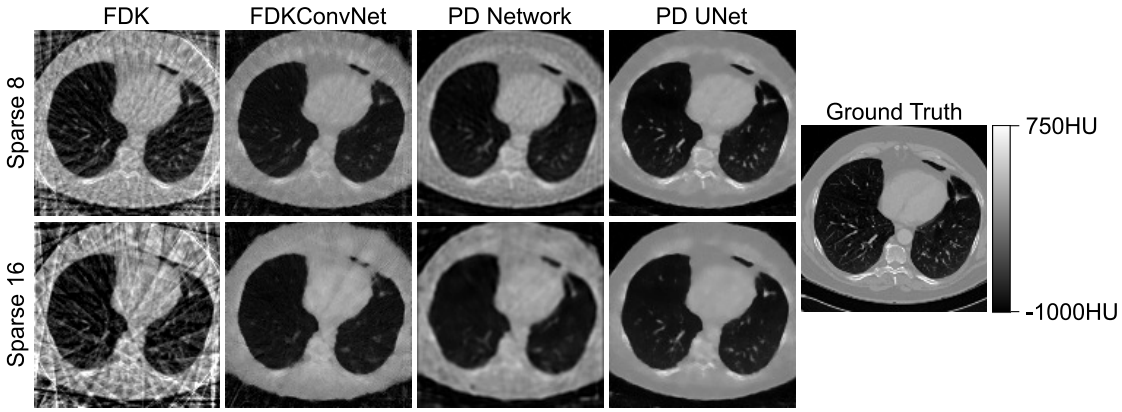


Figure 5.10: Exemplary axial slice from different models/methods.

For comparability, the data normalization, the L_1 loss function, the Adam optimizer (lr=1e-3, β_1=0.9, β_2=0.999) and the number of epochs (151) were kept the same. The effective batch size was set to 16. Training data was simulated by downsampling LungCT-Diagnosis [GBS+15] volumes (42/9/10 for training/validation/test) to cubes with side lengths of 128px = 128mm due to memory limitations. Random flips, rotations and scalings of the volumes were used as augmentation during training. Sparse views were simulated by retaining every 8th or 16th of 360 equiangular projections (called *Sparse 8* or *Sparse 16*, respectively).

5.3.2 Results

Tab. 5.3.2 shows the results of the different models evaluated on the test set. All models outperform the direct sparse view FDK reconstruction by a large margin, while the Primal-Dual models further increase the quality compared to FDKConvNet [JMF+17]. The proposed Primal-Dual UNet results in the lowest errors. Wilcoxon signed-rank tests reveal that the proposed model significantly outperforms any other model/method pair-wise (p-value $< 0.5\%$).

Fig. 5.3.2 shows an exemplary axial slice from the different models for *Sparse 8* (top row) and *Sparse 16* (bottom row). FDKConvNet does not seem to have learned anatomical structures and merely attempts to suppress streaking artifacts. Primal-Dual Network produces results that look blurrier and noisier than FDKConvNet's outputs but

anatomical structures, e.g. costal cartilage, are preserved better. The reconstructions of Primal-Dual UNet are superior compared to Primal-Dual Network. Tissues with high attenuation coefficients are clearly distinguishable from soft tissues and edges are well preserved, e.g. vertebrae, even for the higher sparsity factor *Sparse 16*.

5.3.3 Discussion and Conclusion

The proposed Primal-Dual UNet for cone beam reconstruction not only outperforms other methods – Primal-Dual Network in particular – in quality but also in memory requirements and is more than twice as fast during both training and inference while retaining data consistency wrt. the cone beam projections, as opposed to FDKConvNet. Moreover, the training of the proposed network is much more stable compared to Primal-Dual Network. However, the main limitation is still the memory consumption: with enabled mixed precision, the inference takes ∼9GB of GPU RAM for even these unrealistically low resolution volumes and projections and a batch size of 1. Training consumes even more space: a *Sparse 4* version of Primal-Dual Network did not even fit into the 48GB of an Nvidia RTX A6000.

Since usually, not the entire volume needs to be reconstructed during an intervention, future work will focus on reducing the memory requirements by only reconstructing volumes of interest. Moreover, this preliminary work is based on simulations and has to be evaluated for real cone beam CT data. The Pytorch implementation is available on Github[3].

5.4 Algebraic Prior Knowledge: Cone Beam Projection Interpolation for Circular Trajectories

This section is mainly based on the publication "Trajectory Upsampling for Sparse Conebeam Projections using Convolutional Neural Networks" [ERH+21].

The prior knowledge that is primarily used for this method is Algebraic Prior Knowledge *and secondarily* Deep Learning Prior Knowledge*.*

5.4.1 Analytical Projection Interpolation

As described in [NHD+07], cone beam projections can be approximately interpolated by using (Eq. 24 in [NHD+07])

$$\begin{aligned} g(\lambda + \epsilon\Delta\lambda, \underline{\alpha}) \simeq (1-\epsilon) g(\lambda, \underline{b}(\lambda + \epsilon\Delta\lambda, \underline{\alpha}) - \underline{a}(\lambda)) \\ + \epsilon g(\lambda + \epsilon\Delta\lambda, \underline{b}(\lambda + \epsilon\Delta\lambda, \underline{\alpha}) - \underline{a}(\lambda + \Delta\lambda)) \end{aligned} \tag{5.3}$$

for projections $g(\lambda, \underline{\alpha})$ from source positions $\underline{a}(\lambda)$ in directions $\underline{\alpha}$ and points of interest $\underline{b}(\lambda, \underline{\alpha})$ that are closest to the rotation axis on the line through $\underline{a}(\lambda)$ with direction $\underline{\alpha}$. Unlike [NHD+07], the directions $\underline{\alpha}$ here are chosen to coincide with the projection

[3] `https://github.com/phernst/pd-unet-conebeam`

lines of the projection to be interpolated. This only requires interpolating on the given projections.

5.4.2 CNN Approach

Assuming an equiangular sampling of cone beam projections along a circular trajectory, the presented approach upsamples along the trajectory by subsequently interpolating projections angularly centered between neighboring projections. Simple algorithms like linear interpolation are not applicable because of the sinusoidal structure and perspective distortions caused by the cone beam. A U-Net [RFB15a] is used to approximate this highly complex interpolation because of its large receptive field that is able to capture and trace larger translations in the projections compared to flat CNN architectures. (1) Networks are trained to predict the projection angularly centered between two projections from only its direct neighbors for 2°, 4° and 8° of angular distance (referred to as `nn2`). (2) The number of neighboring input projections is increased from 2 to 4 and 8 neighbors to provide more angular information (referred to as `nn4`, `nn8`). (3) Instead of increasing the number of neighboring projections, the analytical interpolation described in Sec. 5.4.1 with $\epsilon = 0.5$ is used as an additional input which is supposed to guide the network closer to the true interpolation (referred to as `nn2+ana`).

5.4.3 Data Sets and Training

The data of 22 subjects from the CT Lymph Nodes collection [RLS+15] of The Cancer Imaging Archive [CVS+13] is used, consisting of reconstructed volumes of the abdomen with different in-plane spacings that serve as ground truth. Cone beam projections were generated using the CTL toolkit [PFB+19] equiangularly along a circular trajectory with a source to detector distance (SDD) of 1000 mm and a source to isocenter distance (SID) of 750 mm. The flat panel detector consists of 256×256 elements with a pixel size of $4\,\mathrm{mm}^2$ (cone angle of 54.2°). The values were chosen such that most projections were not truncated and to enable a faster training.

The U-Net [RFB15a] has a depth of 5 and is slightly modified. The encoder doubles the number of layers after each average pooling, whereas the decoder halves the number of layers after each nearest neighbor upsampling. The optimizer is SGD with a weight decay of 1×10^{-4} and a learning rate of 6×10^{-3} that gradually drops to 1×10^{-6} by a factor of 0.8 after every 10 epochs of no improvement in validation loss. Every network was trained for 300 epochs using mean squared error (MSE) and another 300 epochs using equally weighted l_1 and MS-SSIM loss similar to [ZGF+17] to focus more on general structures and edges. 16, 4 and 2 data sets were used for training, validation and testing, respectively. For faster convergence, the projections were normalized between 0 and approximately 1 by dividing by the 99th percentile of all projections of all data sets.

Up	Method	NMSE ($\times 10^{-5}$)	PSNR [dB]	SSIM [%]
x2	ana	**10.88±9.10**	**49.13±2.85**	**99.68±0.24**
x2	nn2	17.17±14.21	46.87±2.23	99.40±0.23
x2	nn4	21.55±10.99	45.58±1.94	99.16±0.26
x2	nn8	17.33±8.62	46.45±2.08	99.35±0.25
x2	nn2+ana	10.92±6.92	48.64±2.43	99.61±0.18
x4	ana	32.60±26.17	44.34±3.06	99.03±0.67
x4	nn2	24.02±11.27	45.07±1.96	99.12±0.25
x4	nn4	26.06±10.42	44.61±1.80	98.98±0.27
x4	nn8	23.19±9.27	45.10±1.93	99.13±0.26
x4	nn2+ana	**18.25±9.34**	**46.33±2.32**	**99.32±0.27**
x8	ana	93.52±69.93	39.96±3.67	97.34±1.71
x8	nn2	41.49±22.49	42.88±2.48	98.55±0.59
x8	nn4	43.00±22.86	42.67±2.39	98.43±0.60
x8	nn8	40.75±22.43	42.93±2.52	98.57±0.60
x8	nn2+ana	**32.46±20.36**	**44.07±2.80**	**98.85±0.57**

Table 5.4: Projection errors for different upsampling methods (only calculated on interpolated projections, i.e. excluding the available ground truth projections) ± standard deviation. Bold values indicate the lowest errors for the respective upsampling factor.

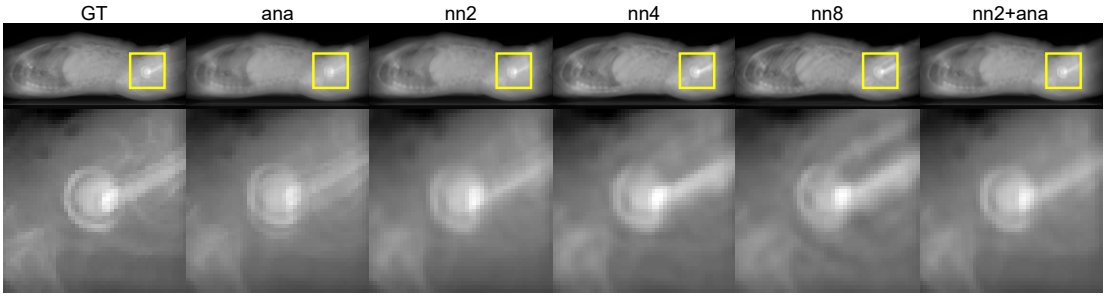


Figure 5.11: Top: Interpolated projections (central part) of x8 upsampling of different interpolation methods compared to ground truth projection (GT). Bottom: Zoomed patches around a hip implant.

5.4.4 Projections

The different interpolation methods are evaluated on the projections first. Except for the analytical upsampling described in Sec. 5.4.1, all methods interpolate the projection angularly centered between the input projections, which is repeated for x4 and x8 upsampling using the corresponding trained networks. For the analytical upsampling, the parameter ϵ is chosen to directly resemble the positions of the projections to be interpolated. Tab. 5.4 shows the results for the error metrics NMSE, PSNR and SSIM averaged over all projections. The calculation of the metrics obviously excludes the non-interpolated projections. Interestingly, the results are quite different for the different upsampling stages.

For the single interpolation (x2, angular difference of 2°), `nn2` gives the best results for NMSE and PSNR. The analytical interpolation however results in the highest SSIM.

Interpolating twice (x4, angular difference of 4°) is done best by `nn2`, this time for all metrics.

Finally, the optimal method for carrying out the interpolation three times (x8, angular difference of 8°) is using `nn2+ana`.

A patch of an exemplary x8 interpolation created with the different methods is shown in Fig. 5.11. Compared to the ground truth patch, the other patches are more blurry. The patch created with the analytical interpolation looks like the superimposition of two projections. The `nn4` and `nn8` patches seem to have more high frequencies than `nn2` and consequently look less blurry. `nn2+ana` is visually closest to the ground truth and the least blurred.

5.4.5 Reconstructions

Evaluating in projection space only does not fully show the benefits of the proposed method. It is also necessary to compare the reconstructions. We decided for the commonly used FDK [FDK84] algorithm as well as ART [GBH70] without interpolated projections initialized with the FDK reconstruction using all interpolated projections.

Method	NMSE [%]	PSNR [dB]	SSIM [%]
full	4.25±7.65	24.75±4.70	72.64±7.56
sparse	11.13±6.70	17.53±1.77	37.64±6.22
ana	6.70±7.80	20.63±3.02	59.74±5.67
nn2	5.55±7.81	22.24±3.64	64.68±6.49
nn4	5.54±7.81	22.27±3.65	64.38±6.39
nn8	6.43±7.69	20.87±3.10	57.62±5.12
nn2+ana	**5.32±7.80**	**22.60±3.78**	**66.50±6.80**

Table 5.5: Reconstruction errors of FDK reconstructions for different upsampling methods from 45 available projections ± standard deviation.

All reconstructions are created with the CTL toolkit [PFB+19]. The ART reconstructions run for 5 iterations with enabled positivity constraint.

Since the number of projections is still relatively small and the resolution of the detector is quite low, the reconstructions will also be compared to the FDK reconstruction using all 360 projections to find lower or upper bounds for the error metrics.

As described previously, though not depending on any reconstruction algorithm, the interpolated projections are supposed to increase the quality of the reconstructions by providing a more appropriate sampling of projections.

This hypothesis is evaluated using the FDK reconstruction algorithm, first. For brevity, only the reconstructions of the highest upsampling (x8, 45 available projections) are investigated. Tab. 5.5 shows the error metrics for the different methods averaged over all axial slices. For reference, the first two rows serve as lower/upper bounds: values for the full FDK describe the errors between the ground truth volume and the volume reconstructed from 360 projections, whereas values for the sparse FDK describe the errors between ground truth and reconstruction from 45 projections. All interpolation methods optimize the sparse FDK reconstruction and are quantitatively closer to the full FDK. `nn2+ana` works best, followed by `nn2`, `nn4`, `nn8` and using only the analytical interpolation. This closely resembles the errors on the projections described in the previous section.

The left column of Fig. 5.12 shows exemplary FDK reconstructions using the different methods. Compared to the direct FDK reconstruction from 45 projections (`sparse`), every method reduces the streak artifacts. The analytical upsampling (`ana`), however, basically results in a radially blurred reconstruction. None of the CNN-based reconstructions suffers from streak artifacts or radial blur, but they appear slightly more blurred than the `sparse` FDK reconstruction. As expected from the quantitative analysis, `nn2+ana` also creates the best visual result.

ART provides another simple reconstruction algorithm. Due to its iterative nature, it is inherently slower than FDK but enables simply adding additional constraints resulting in reconstructions of higher quality. For a better convergence, ART is initialized with

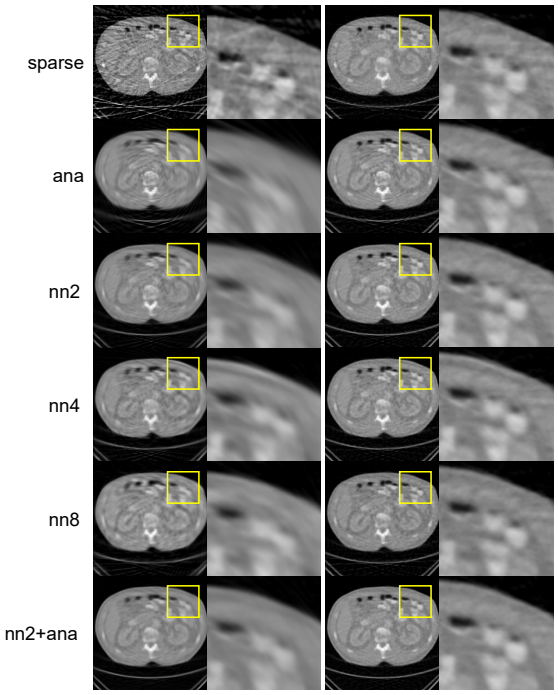


Figure 5.12: Reconstructions for different upsampling methods. Left column: FDK. Right column: ART initialized with FDK.

Init.	NMSE [%]	PSNR [dB]	SSIM [%]
zero	2.74±0.97	23.40±0.76	67.21±3.02
sparse	2.31±0.83	24.06±0.76	67.24±2.96
ana	2.19±0.82	24.32±0.80	69.60±3.17
nn2	1.76±0.80	25.48±1.00	72.89±3.32
nn4	1.75±0.80	25.51±1.01	72.93±3.30
nn8	2.12±0.85	24.51±0.85	69.66±3.10
nn2+ana	**1.64±0.78**	**25.79±1.03**	**74.28±3.33**

Table 5.6: Reconstruction errors of ART reconstructions for different upsampling methods from 45 available projections ± standard deviation.

another reconstruction. In our experiments, we use the FDK reconstructions of the different interpolation methods and run ART with only the 45 available projections, which results in the best compromise between reconstruction time and quality. Tab. 5.6 shows the error metrics. Zero-initialized ART and sparse-FDK-initialized ART are shown for reference. In all cases, ART outperforms FDK. Again, `nn2+ana` works best, followed by the other methods in the same order as in the FDK reconstructions.

The right column of Fig. 5.12 shows exemplary ART reconstructions using the different methods. They are not only quantitatively closer to the ground truth but also qualitatively outperform their FDK counterparts. There are only slight visual differences of the ART initialized with the different FDK reconstructions. For the `sparse` case, edges are preserved well but tissues of the same absorption coefficient appear noisy. `nn2+ana` has the best visual quality with the least noise and the best edge preservation compared to the other methods.

Note: The error values reported in the previous tables are different from those reported in the original publication [ERH+21] where especially the SSIM values were much closer to 100 %. This was due to a different dynamic range variable in the calculation of the SSIM, set to 1. Here, however, it was set to 0.0269 for reconstructions, resembling the 99th percentile of the attenuation coefficients, and 7.1230 for projections, resembling the 99th percentile of projection values in the data sets. Not only does this scale the range of SSIM such that differences can be distinguished better during evaluation, it also resembles the meaning, and therefore intended use, of this dynamic range variable closer, excluding a large interval of attenuation values without information (that is, $[0.0269, 1]$). Since the SSIM was also used in the loss function for the trainings with this differently set parameter, the networks were retrained resulting in slightly different values for NMSE and PSNR, as well.

5.4.6 Discussion

Increasing the number of neighboring projections does not increase the quality of the interpolated projections. Since the additional projections are only provided to the CNN as input channels and the convolutions are carried out per channel, it is possible that (without any special weight initialization) the information from more distant neighbors is not local enough to be considered as helpful knowledge during backpropagation. Moreover, increasing the number of input projections even impairs the prediction quality. Further tests need to investigate why different interpolation methods work best for certain upsampling stages.

The simulated projections do not contain noise, are almost not truncated, have a low resolution and a rather large pixel spacing. Further experiments need to focus on more realistic detector and gantry parameters and the method needs to be tested on real data, especially including interventional instruments and other artifact creating influences.

The used error metrics only give a rough impression of the quality. Due to the blurring of edges caused by the interpolation, future work needs to focus on how exactly

mappings of edges are changed as well as how the reconstructions compare to other state-of-the-art methods.

Using the neighboring projections as input channels of the U-Net is a rather straightforward way. As with other deep learning methods, it is conceivable that another network architecture can extract more information from the input data and thus improve the quality even further, which will be part of future experiments. The code is available on Github[4].

5.5 Algebraic Prior Knowledge: Sparse View Deep Differentiated Backprojection

This section is mainly based on the publication "Sparse View Deep Differentiated Backprojection for Circular Trajectories in CBCT" [ERN21].

The prior knowledge that is primarily used in this method is Algebraic Prior Knowledge *and secondarily* Deep Learning Prior Knowledge.

5.5.1 Approach

The errors caused by discretization occur at different stages in the reconstruction algorithm. For this reason, it is necessary to dedicate different networks to these stages and evaluate if combined networks can approximate the Hilbert inversion with errors from different stages more accurately than others.

In total, six networks are trained. (1) For comparison, a post-processing network is trained that enhances the FDK reconstruction of 36 projections for sagittal or coronal slices. (2) A network that enhances the DBP (Eq. 3.3) of 36 projections to approximate the DBP of fully sampled projections. (3) A Hilbert inversion network that inverts fully sampled DBP planes. (4) Like (3) but with an additional FDK reconstructed (360 projections) plane as input. (5) A Hilbert inversion network that inverts DBP planes from 36 projections and enhances them to approximate reconstructions of fully sampled projections. (6) Like (5) but with an additional FDK reconstructed (36 projections) plane as input.

All networks share the same U-Net-like architecture except for the number of input/output channels and are trained on both coronal and sagittal planes-of-interest.

For the final reconstructions, the following combinations of networks are investigated: Network (1) for comparison (`fdkconv`). Network (2) + Network (3) (`s2f_inv`). Network (2) + Network (4) (`s2f_inv3`). Network (5) (`inv_sp`). Network (6) (`inv_sp3`).

5.5.2 Spectral Blending

As described in [HSY20], the reconstructed planes of the different Hilbert directions can be combined using spectral blending in order to minimize the missing frequency information. A bow-tie mask is multiplied with the Fourier transforms of the reconstructed

[4] `https://github.com/phernst/conebeam_interpolation`

Method	NMSE [%]	PSNR [dB]	SSIM [%]
fdkconv	1.63±0.52	23.53±0.91	65.86±3.13
s2f_inv	3.18±2.70	21.86±2.45	66.35±4.87
s2f_inv3	7.65±2.56	16.74±1.00	31.60±3.12
inv_sp	3.17±3.34	21.96±2.31	65.41±4.90
inv_sp3	**1.24±0.46**	**25.08±0.97**	**72.05±3.93**

Table 5.7: Errors w.r.t. ground truth of reconstructions from coronal planes-of-interest averaged over axial planes ± standard deviation.

planes and added. The masks are chosen such that the frequency information from both planes complement each other. By angular blurring of the mask, frequency information that is contained in both planes can be combined, as well.

5.5.3 Data Sets and Training

The data of eleven subjects from the CT Lymph Nodes collection [RLS+15] of The Cancer Imaging Archive [CVS+13] is used, consisting of reconstructed volumes of the abdomen that serve as ground truth. Cone beam projections were generated using the CTL toolkit [PFB+19] equiangularly along a circular trajectory with a source to detector distance (SDD) of 1000 mm and a source to isocenter distance (SID) of 750 mm. The flat panel detector consists of 1024×1024 elements with a pixel size of $1\,\mathrm{mm}^2$ (cone angle of 54.2°).

A slightly modified U-Net [RFB15b] with a depth of 5 is used. The encoder doubles the number of layers after each average pooling, whereas the decoder halves the number of layers after each bilinear upsampling. SGD is used as the optimizer with a weight decay of 1×10^{-4} and a learning rate of 5×10^{-2} that gradually drops to 1×10^{-2} by a factor of 0.8 after every 10 epochs of no improvement in validation loss. Every network was trained for 300 epochs using MSE. Eight subjects were used for training, two for validation and the remaining one for testing. For faster convergence, the reconstructed planes are normalized between 0 and roughly 1 by dividing by the 99th percentile of the attenuation coefficients of all axial planes of all data sets. Similarly, the Hilbert planes are normalized by dividing by the standard deviation of all Hilbert planes of all data sets. Random horizontal flips were used as augmentation during training.

5.5.4 Results

Tab. 5.7 shows the mean errors of axial slices using coronal planes-of-interest for the different combinations of networks as described in Sec. 5.5.1, which include NMSE, PSNR and SSIM. The lowest errors are achieved using `inv_sp3`, followed by the simple post-processing network `fdkconv`. All other combinations result in worse errors, the worst being `s2f_inv3` with an NMSE which is almost seven times higher than the best

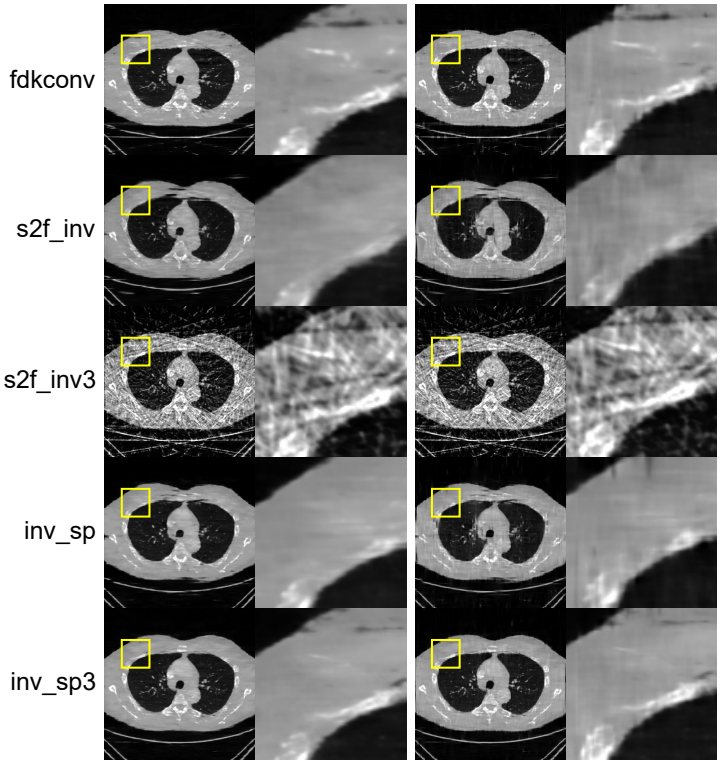


Figure 5.13: Exemplary reconstruction of different methods. Left: using coronal planes-of-interest. Right: after spectral blending.

Method	NMSE [%]	PSNR [dB]	SSIM [%]
fdkconv	1.98±0.72	22.69±1.28	65.23±4.29
s2f_inv	3.53±2.06	20.86±1.40	64.75±3.72
s2f_inv3	8.12±2.55	16.44±0.86	30.14±3.04
inv_sp	4.22±3.14	20.13±1.80	62.84±4.17
inv_sp3	**1.34±0.49**	**24.67±1.07**	**71.82±3.44**

Table 5.8: Errors w.r.t. ground truth of reconstructions from sagittal planes-of-interest averaged over axial planes ± standard deviation.

`inv_sp3`. An important thing to note here is that the additional FDK plane of Network (4) was reconstructed using 360 projections while training, whereas during the inference for the combination with Network (2), only 36 projections were available for the FDK reconstruction and necessarily introduced streaking artifacts. The other combinations `s2f_inv` and `inv_sp` have only slightly worse errors than `fdkconv`. Despite already showing large differences in the error metrics between `fdkconv` and `inv_sp3`, the Wilcoxon signed rank test was performed between the distributions of the two methods which resulted in statistically significant improvements of `inv_sp3` (p-value < 0.01).

The left column of Fig. 5.13 shows an axial slice reconstructed from coronal planes using the different methods. Except for `s2f_inv3`, all combinations result in less discontinuous reconstructions than `fdkconv`. `s2f_inv` seems to smooth out highly absorbing tissues. The best visual appearance for this slice is achieved using `inv_sp` with the least discontinuities and the highest edge preservation. As described earlier, `s2f_inv3` necessarily performs worse because of the way it was trained. However, since the streaking artifacts are very prominent, it can be assumed that Network (4) mainly focuses on the FDK input rather than the DBP plane. Again, the Wilcoxon signed rank test was performed between `fdkconv` and `inv_sp3`, again resulting in statistically significant improvements of `inv_sp3` (p-value < 0.01).

The same behavior as in Tab. 5.7 can be seen in Tab. 5.8, but here for sagittal planes-of-interest. Interestingly, all errors are slightly worse than their counterpart on coronal planes-of-interest except for `fdkconv`. For brevity, the qualitative results are not shown.

5.5.5 Spectral Blending

The blurring radius of the bow-tie mask for the spectral blending of reconstructions from coronal and sagittal planes-of-interest seems to be an essential parameter for the quality of the final reconstructions, as prior tests have shown. The influence of different blurring radii is shown in Fig. 5.14 for `inv_sp3`. There seems to be an almost linear dependency between the radius and the different error metrics: the higher the radius, the closer the reconstruction to the ground truth. This is reasonable because more and more frequencies from both planes are accounted for when increasing the radius. For this reason, the blurring radius is set to 90°.

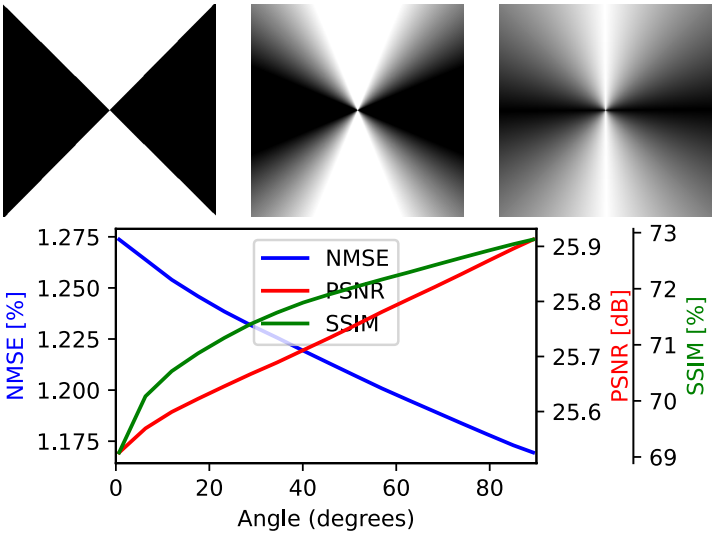


Figure 5.14: Top: Masks without (left), 45° (center) and 90° blurring (right). Bottom: Error metrics for `inv_sp3` reconstructions after spectral blending depending on blurring radius of masks.

Method	NMSE [%]	PSNR [dB]	SSIM [%]
fdkconv	1.45±0.44	24.12±0.87	67.69±3.91
s2f_inv	2.07±1.11	23.23±1.40	54.59±7.54
s2f_inv3	7.92±2.54	16.57±0.91	30.32±3.10
inv_sp	2.35±1.85	22.81±1.78	60.30±5.40
inv_sp3	**1.05±0.37**	**25.91±0.88**	**73.01±3.27**

Table 5.9: Errors after spectral blending ± standard deviation. `sparse_fdk` shows the errors of an FDK reconstruction from 36 projections for reference.

Tab. 5.9 shows the errors of the different methods after spectral blending. Compared to the reconstructions without spectral blending, the errors are even lower. For the best method `inv_sp3`, the SSIM is increased by 0.96 % and 1.19 % compared to coronal and sagittal plane-of-interest reconstructions. A final Wilcoxon signed rank test was performed for these spectrally blended results which again showed statistically significant improvements (p-value < 0.01) and therefore the superiority of the method `inv_sp3` wrt. `fdkconv`.

The right column of Fig. 5.13 shows the reconstructions after spectral blurring. Almost all methods benefit from the additional sagittal information. Visual differences cannot be observed for `s2f_inv3` because of the focus of Network (4) on the FDK input which does not incorporate different information for sagittal or coronal planes. The reconstruction using `inv_sp` introduces some additional discontinuities, probably caused by the worse quality of the network on sagittal planes.

Note: As in the previously described method, the values of the error metrics in above tables differ from those in the original work [ERN21]. Again, the dynamic range parameter of SSIM was set differently, which would make the comparison of the different methods less meaningful. Therefore, the SSIM values were recalculated with the range parameter set to 0.0269, resembling the 99th percentile of the attenuation coefficients of all data sets. Furthermore, an error in the calculation of the PSNR was detected (missing a square root for the normalization factor in the calculation of the RMSE, resulting in a wrongly scaled RMSE). The results of Sec. B.2 were used to fix these errors by subtracting a constant value from the wrong PSNR values. This implies that differences between PSNR values of different methods remain unchanged.

5.5.6 Discussion

In general, only one of the proposed combinations in fact improves the simple post-processing baseline `fdkconv`, which is `inv_sp3`. A possible explanation for this is that, compared to the other combinations, `inv_sp3` can directly learn to extract the most useful information from both the sparse view FDK (which already contains correct frequency information) and DBP (which is able to incorporate information of truncated projections as well as different information from sparse views compared to FDK). The other combinations do not include the FDK reconstruction (`inv_sp` and `s2f_inv`) or are not trained end-to-end (`s2f_inv` and `s2f_inv3`), which does not allow the gradients to flow back completely.

As described earlier, the additional input of Network (4) was the FDK reconstruction of 360 projections while training and of 36 projections while testing `s2f_inv3`. For further tests, the output of `fdkconv` could be used as this additional input to be closer to what the network was trained on.

Moreover, there is a significant difference in accuracy of all networks between coronal and sagittal planes, which might be caused by less variance in the sagittal planes. Additional data or different augmentation techniques could resolve this.

The spectral blending results in even lower errors but depends on the masks that are used. The almost linear dependency of the blurring radius of the mask on the final error metrics (Fig. 5.14) suggests increasing the radius even further or using masks of non-bow-tie shape.

We discovered that all networks including some kind of Hilbert inversion need high learning rates $\geq 10^{-2}$. Setting them lower resulted in both higher loss values and less robust trainings, which seems rather counter-intuitive. Trainings with learning rates between 10^{-4} and 10^{-5} (cf. [HSY20]) did not converge at all, which might be related to the data set or normalization of the data.

To further increase the reconstruction quality, it is conceivable to additionally input neighboring planes-of-interest to the networks to gain more spatial information. In addition, the effect of choosing different values for ϵ for creating the partial derivatives was not investigated and needs further tests. The code is available on Github[5].

5.6 Temporal and Model Prior Knowledge: Interventional Instrument Enhancement in C-arm Reconstructions from Few Projections using Prior Scans

This section is mainly based on the publication "Towards Patient Specific Reconstruction Using Perception-Aware CNN and Planning CT as Prior" [GER+22] © 2022 IEEE.

The prior knowledge that is primarily used in this method is Temporal and Model Prior Knowledge *and secondarily* Deep Learning Prior Knowledge.

In this work, we propose a CNN-based method that utilizes both a planning CT and an interventional CT of the same subject to produce artifact-reduced and data-consistent (wrt. CBCT intensities) reconstructions. The deep learning models designed for the reconstruction task typically use the pixelwise MSE to measure the reconstruction error. However, MSE is considered an unreliable metric in image quality assessment studies, as it does not capture the structural relationship in a pixel neighborhood like humans do [WB09]. Therefore, we propose a perception-aware loss function, which helps the model capture beyond pixelwise differences.

5.6.1 Architecture: Dual Branch Prior-Net

The proposed CNN architecture is portrayed in Fig. 5.15.

Two inputs are provided to the network. One of them is the high-quality planning CT of the subject, acquired prior to the surgery. The other one is the sparsely sampled interventional CT of the same subject. Only the interventional image contains the surgical instrument. The network is logically divided into two components, Extraction and Fusion.

[5] `https://github.com/phernst/sparse_dbp`

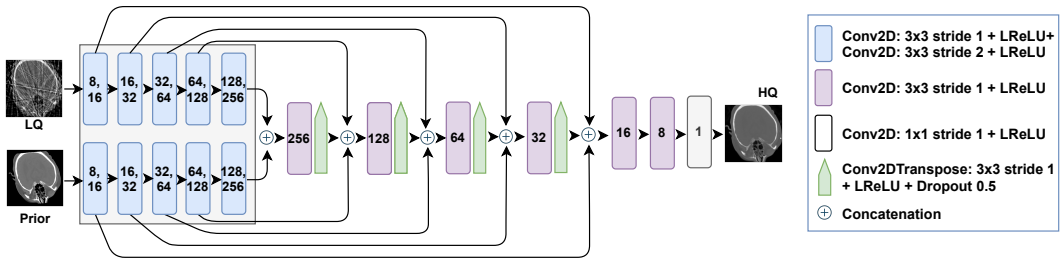


Figure 5.15: The proposed architecture. LQ: sparse-sampled FDK reconstructed interventional CBCT. Prior: planning CT (conventional CT scanner). HQ: ground truth (fully-sampled FDK-reconstructed interventional CBCT). Filter numbers of Conv2D shown inside the boxes. © 2022 IEEE

In the Extraction, both inputs are downsampled to increase the receptive field of the network. Although the planning and interventional images have different intensities for the same pixel of the same subject, we hypothesize that the representations captured from the same subject's structural details in the Extraction facilitate the network to produce data-consistent (wrt. CBCT intensities) interventional reconstructions. Each convolutional block in the Extraction consists of 2 convolutions, each followed by Leaky ReLU [MHN+13] activation to overcome the 'dead-neuron problem' of ReLU. The last convolution of the block uses strides of 2 in order to downsample the image, instead of a pooling operation [SMB10], such that the network remains sensitive to the location of a structure in the image.

In the Fusion, high resolution information from the prior CT and interventional CBCT are combined and passed to the same level of the Fusion through skip-connections. In this component, a convolution is followed by a transposed convolution with 3×3 kernels and strides of 2 for upsampling. The last layer is a 1×1 convolution followed by Leaky ReLU, which produces the final reconstruction.

5.6.2 Loss Function

SSIM has been used in image and video quality assessment tasks over the years, as it resembles how humans perceive differences between two images [WT97]. It is calculated as a weighted product of the three measures: luminance (L), contrast (C) and structure (S) (see Sec. 3.8.4). However, SSIM does not penalize enough when a small disconnected structure like a needle is not reconstructed. Therefore, multi-scale SSIM (MSSIM) [WSB03] is more appropriate for our task, which considers the structural details at various resolutions. In MSSIM, the image is downsampled by a factor of two iteratively. The *contrast* and *structure* comparison is done at all scale levels, as shown in Eq. 5.4, where $\mathbf{x}$ and $\mathbf{y}$ are two images, α, β_s and γ_s are the weights for *luminance*, *contrast* and *structure* at scale s, respectively. The *luminance* is calculated only for the highest scale N, as the human eye is more perceptive to changes in texture or edges

than smooth regions. Eq. 5.5 portrays the proposed perception-aware loss, considering $(1 - MSSIM)$ as the structural dissimilarity.

$$MSSIM(\mathbf{x}, \mathbf{y}) = [L(\mathbf{x}, \mathbf{y})]^{\alpha} \cdot \prod_{s=1}^{N} [C_s(\mathbf{x}, \mathbf{y})]^{\beta_s} [S_s(\mathbf{x}, \mathbf{y})]^{\gamma_s} \tag{5.4}$$

$$Loss_{MSSIM}(\mathbf{x}, \mathbf{y}) = (1 - MSSIM(\mathbf{x}, \mathbf{y})) + MSE(\mathbf{x}, \mathbf{y}) \tag{5.5}$$

5.6.3 Data Set and Preprocessing

Fifty head CT volumes from the publicly available Mayo Clinic data set [MCH+20] were used for the task. They were acquired helically and served as the planning CT. The needle data set comprises three scans of an ablation needle by NeuWave Medical inserted into an abdominal phantom in different positions, scanned in a KIDS-CT system. The interventional scans had to be simulated, as the head and needle scans were available separately. To this end, cone-beam projections of each data set were created using torch-radon [Ron20] with a C-arm span of 1200 mm, 620×480 detector pixels with spacing of 0.616 mm and a circular trajectory with 360 equi-angular projections. Prior to projecting the needle volumes, the surrounding phantom was removed by thresholding. Assuming the cone-beam projections to be ray integrals, the needle and head projections were combined by summation. This provided the cone-beam projections to reconstruct the interventional scans. The sparsely sampled interventional volumes (LQ) were reconstructed from 13 simulated projections as in [GKR+21] and the ground truth (HQ) from all the projections using FDK. Reconstructing these simulated projections (subsampling them angularly to simulate sparse views) results in volumes similar to interventional scans including the described artifacts.

5.6.4 Training Details

The models were trained on the axial slices of the CT volumes. Each slice was cropped to 384×384, to retain only the useful information. To speed up convergence of the network, the intensities were normalized to roughly the unit interval by converting Hounsfield units to mass attenuation coefficients and dividing by the 99th percentile of all training data's mass attenuation coefficients. The data set was split into training (17180), validation (11631) and test (10296) slices. We applied augmentation during training in terms of axial rotations of the needle, by rolling the list of needle projections by a random amount before reconstruction. Also, rotation, scaling and left-right flips were applied to the reconstructed slices. Additional in-plane rotation in the range $[-10°, 10°]$ was applied to the planning CT, to misalign it wrt. the interventional CBCT. The models were trained using ADAM optimizer (learning rate$=1 \times 10^{-4}$). Each experiment was run with early-stopping (after 20 epochs of no improvement) on the validation set. The model with the highest validation PSNR was selected and used for the evaluation on test data.

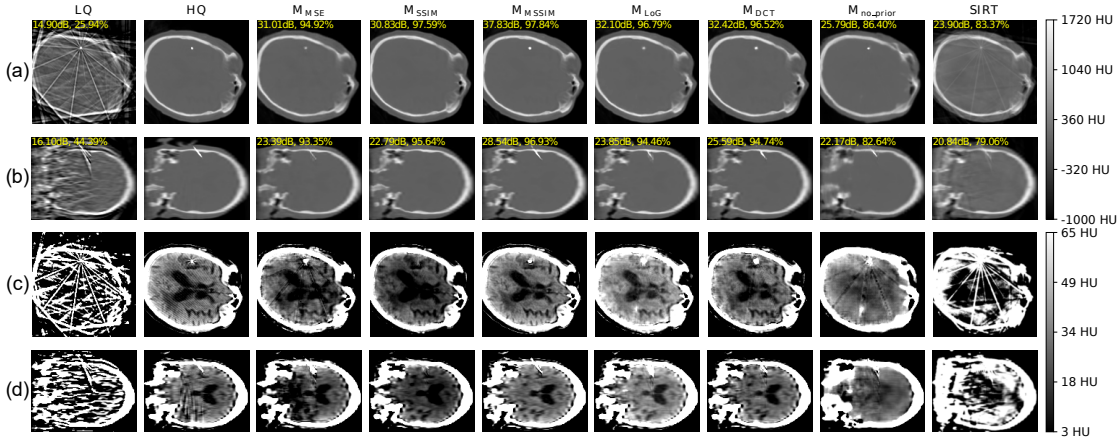


Figure 5.16: Reconstructions of one test subject. Rows (a), (b): axial and coronal slice with $[-1000, 1720]$ HU to compare the needle's reconstruction quality. Rows (c), (d): same slices with $[3, 65]$ HU to compare the quality of reconstructed soft-tissue. LQ: sparse-sampled input. HQ: corresponding ground truth. Remaining columns: predictions/reconstructions of different methods. PSNR and SSIM in yellow. © 2022 IEEE

Two baselines were considered for the proposed deep learning network. For the first one, M_{no_prior}, the sparse-sampled input was again passed instead of the planning CT and uses pixelwise MSE loss. The other one, SIRT, implemented in ASTRA toolbox [vAPC+16], used the planning scan as initialization. We trained four models using planning CT and different losses. The model M_{MSE} used pixelwise MSE as the loss. The model M_{MSSIM} used the proposed perception-aware loss L_{MSSIM}, and M_{SSIM} used SSIM instead of MSSIM. The last two models were trained with perception-aware losses proposed in [GKR+21], where M_{LoG} represents the model trained using loss calculated in Laplacian of Gaussian space and M_{DCT} denotes the model using loss calculated in discrete cosine transform (DCT) space. These two models served as baselines for the proposed perception-aware loss.

5.6.5 Quantitative Results

We evaluate the performance of the models using the following metrics: MSE, PSNR and SSIM. Tab. 5.10 portrays that M_{MSSIM} performed the best with respect to all metrics. The non-deep-learning baseline SIRT improved the FDK reconstructions, but performed worse and showed higher standard deviation compared to the deep learning models. We also observed that the deep learning baseline model not equipped with planning CT (M_{no_prior}) performed significantly worse than M_{MSE} (p-value of 1×10^{-5} in a one-sided paired t-test). This supports our hypothesis that the representations from the planning CT helped the network to produce better reconstructions. The planning CT model using the proposed loss (M_{MSSIM}) improved the reconstruction quality over the one using MSE loss (M_{MSE}) significantly (p-value of 1.2×10^{-4} in a one-sided paired

Table 5.10: Quantitative evaluation of reconstruction quality of deep learning models, SIRT and FDK. Mean and standard deviation over all axial 2D slices of the test set. Best results marked bold.

Model	SSIM [%]	PSNR [dB]	MSE $(\times 10^{-4})$
FDK	31.4 ± 13.1	17.64 ± 5.05	241.00 ± 56.00
SIRT	61.1 ± 14.0	21.02 ± 8.98	190.00 ± 40.00
$M_{\text{no\_prior}}$	80.5 ± 7.4	27.46 ± 5.05	27.50 ± 2.00
M_{MSE}	92.2 ± 3.5	32.89 ± 5.32	8.65 ± 2.60
M_{MSSIM}	$\mathbf{95.6 \pm 2.7}$	$\mathbf{36.14 \pm 6.05}$	$\mathbf{4.99 \pm 2.20}$
M_{SSIM}	95.2 ± 2.4	33.42 ± 5.42	7.89 ± 3.40
M_{LoG}	94.4 ± 2.7	34.10 ± 5.18	6.47 ± 2.40
M_{DCT}	94.3 ± 2.8	34.14 ± 5.57	6.70 ± 1.90

t-test). We also observe that all the planning CT models using surrogate losses produced lower MSE than the one trained with only MSE loss.

5.6.6 Qualitative Results

Fig. 5.16 provides qualitative comparison of the reconstructions produced by deep learning models and the non-deep-learning baseline SIRT. We chose to display the results using two different windowings, as one of them helps in comparing the needle's reconstruction quality, and the other in comparing the reconstruction of detailed brain structures and soft-tissue. For most of the cases, the model M_{MSSIM} reconstructs the needle precisely compared to the others, as seen in Fig. 5.16 (a-b) and 5.17. However, the two perception-aware models M_{DCT} and M_{LoG} performed similarly to M_{MSSIM} in many cases. For some cases, the needle produced by M_{DCT} and M_{LoG} were broader or wider in diameter, as seen in Fig. 5.16 (a). Further, we see that all the models could reconstruct the needle except M_{SSIM}. This supports our hypothesis of using multi-scale SSIM instead of SSIM in the loss function, as SSIM could not help the model capture the small disconnected structures. Interestingly, only M_{MSSIM} seems to be able to correctly reconstruct the mean intensity of the brain tissue and additionally preserve high frequency content. In other models, either some edge information gets lost or the brain tissue is hyper- or hypointense on average. Fig. 5.16 (c) shows that M_{MSE} and SIRT could not remove the streaking artifacts completely. Also, the model without planning CT M_{no_prior} and SIRT produced reconstructions having poor soft-tissue contrast, as seen in Fig. 5.16(c-d). We can also see that the sparse-sampled interventional scan had minimal information about the soft-tissue or other structural details of the head. This indicates that the incorporation of the same subject's planning CT helped the network to reconstruct the high intensity structures (bone and teeth) and the soft-tissue better. The code is publicly available on Github[6].

[6] `https://github.com/suhitaghosh10/interventional-CT`

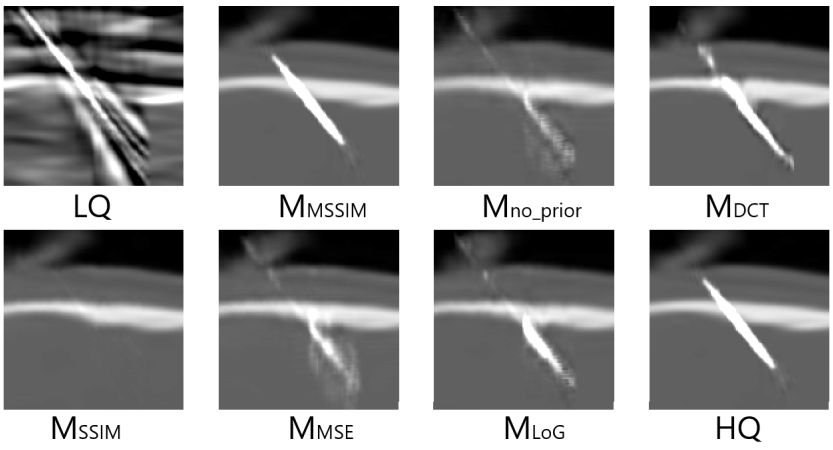


Figure 5.17: Region of interest around the needle produced by the models for the LQ shown in Fig. 5.16 (b). © 2022 IEEE

5.7 Temporal and Model Prior Knowledge: Segmentation as Auxiliary Task to Guide Sparse View CBCT Reconstruction Incorporating Prior Scans

This section is mainly based on the publication "Dual Branch Prior-SegNet: CNN for Interventional CBCT using Planning Scan and Auxiliary Segmentation Loss" [EGR+22].

The prior knowledge that is mainly used in this method is Temporal and Model Prior Knowledge.

The main contributions of this work are:

1. evaluating the performance of the Dual Branch Prior-Net with an additional segmentation head guiding the reconstruction task and

2. determining the limits for the misalignment of the prior scan for in-plane rotations.

5.7.1 Architecture: Dual Branch Prior-SegNet

The deep learning architecture in this work is based on the network in Sec. 5.6 [GER+22], which is a multi-scale dual branch CNN extracting features from both a sparse view interventional CBCT scan and a high resolution planning scan (used as the prior) separately. These are combined via UNet-like skip connections in the decoding path. After the last upsampling, we add another convolutional block parallel to the reconstruction block of the original network which is supposed to segment interventional instruments. The hypothesis is that giving the network the additional task of segmentation increases the quality of the reconstruction since it is forced to learn what causes the most prominent streaking artifacts.

Table 5.11: Parameters of the simulated C-arm CT geometry.

Parameter	Value
Detector binning	4 x 4 pixels
Detector columns x rows	2480 x 1920
Effective detector pixel size	616 µm
Source-to-detector distance	1200 mm
Source-to-isocenter distance	800 mm
Protocol	Axial scan

5.7.2 Loss Function

The loss function used for training is a linear combination of the reconstruction loss and a segmentation loss:

$$L(p, g) = \lambda_1 \cdot L_{MSE}(p_r, g_r) + \lambda_2 \cdot L_{Dice}(p_s, g_s)$$

for the prediction p and the ground truth g, the subscripts r and s denoting the reconstruction and segmentation, L_{Dice} being the Dice loss and $\lambda_1,\ \lambda_2$ scaling factor which have to be set empirically (here: $\lambda_1 = 1,\ \lambda_2 = 1\text{e}{-3}$).

Note that the original Dual Branch Prior-Net [GER+22] was not robust regarding the loss function: except for a combination of MSE loss and Multiscale SSIM loss, all models trained with different loss functions resulted in hyper- or hypointense reconstructions wrt. attenuation coefficients, were not able to remove the streaking artifacts or did not reconstruct the interventional instruments at all.

In this setting, however, no big differences in performance could be found. For this reason, the simple MSE loss was chosen as reconstruction loss.

5.7.3 Data Set and Preprocessing

The training data was created from the LungCT-Diagnosis data set (see Sec. 5.1.4), as opposed to the Mayo Clinic heads data set (see Sec. 5.1.3) used by [GER+22], and in-house needle scans from the NeuWave Medical Needles data set (see Sec. 5.1.2). The interventional data was simulated using torch-radon [Ron20] by superimposing and then reconstructing the projections of the lungs and needles. The parameters of the CT acquisition geometry are shown in Tab. 5.7.3.

Due to the rather limited field of view of the simulated C-arm geometry, many projections were truncated which resulted in even more pronounced artifacts in the reconstructions. This poses an additional challenge to the network training. However, dose reduction for interventional CBCT scans of lungs is more important than for cranial scans. Due to the severity of diseases that are treated with ablations inside the head, e.g. brain tumors, the relatively high amount of X-radiation during full dose scans is still not as fatal as the diseases themselves, which is why dose reduction usually plays a minor role for head scans.

Table 5.12: Mean and standard deviation over all axial 2D slices of the test set without misalignment. *(no mis)* indicates models trained without misalignment of the prior which are excluded for the reasons stated in the description.

Model/Method	SSIM [%]	PSNR [dB]	RMSE [HU]
FDK	17.47±7.35	10.33±1.48	1473.60±349.98
FDKConvNet	64.16±11.18	21.78±2.38	234.40±67.38
Dual Branch Prior-Net (no mis)	**97.90±1.41**	38.81±2.40	34.84±14.74
Dual Branch Prior-SegNet (no mis)	97.04±2.04	38.91±3.80	40.92±34.13
Dual Branch Prior-Net	96.71±3.31	41.09±3.50	28.70±15.09
Dual Branch Prior-SegNet	**97.15±3.47**	**43.97±4.95**	**23.21±18.11**

Opposing to this, reducing dose for lung scans is usually desirable. This is because abnormal tissues treated with ablations inside the lungs are usually less fatal than diseases that X-radiation of a (or potentially several) full dose scan(s) could induce. Moreover, treatments in the lung do not have to be as precise as in the brain and are performed much more frequently than brain ablations.

The segmentations were created by thresholding the needle scans at 900 HU. All axial slices were normalized by converting to attenuation coefficients and dividing by the 99th percentile of all values of the data set.

5.7.4 Training Details

The networks were trained for 150 epochs using Adam (lr=1e-3) with a batch size of 32 and mixed precision. Online augmentations were performed including random rotations, scalings and flips and for some trainings up to ±5deg in-plane misalignment of the prior scan.

5.7.5 Results

Tab. 5.7.5 shows the results of the different models/methods evaluated on the test set (no misalignment). All models outperform the direct sparse view FDK reconstruction by a large margin, while the Dual Branch models further increase the quality noticeably compared to the post-processing FDKConvNet (FBPConvNet [JMF+17] trained on axial slices of FDK reconstructions). The models trained with augmentations and misalignments do not clearly increase the SSIM (or, in case of Prior-Net, even decrease the SSIM) but have a positive influence on both the PSNR and RMSE values.

Since the *(no mis)* models are strongly dependent on an exact registration of interventional and planning scan, as will be shown in the following paragraphs and in Fig. 5.7.5 b and c, they are less suitable for an application in medical intervention and, therefore, will be excluded from the following tests.

The proposed model, Dual Branch Prior-SegNet, results in the lowest errors. Wilcoxon signed-rank tests (excluding the *(no mis)* models) reveal that the proposed model sig-

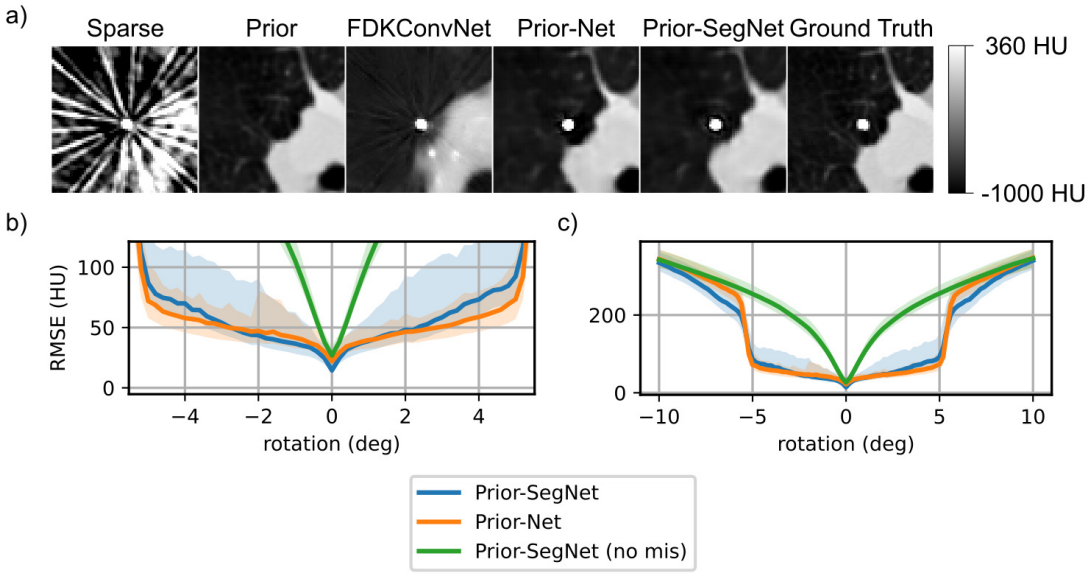


Figure 5.18: a) Exemplary ROI around needle of different models/methods. b, c) Reconstruction errors (RMSE) using misaligned (rotated) prior (interquartile range: semi-transparent, median: solid line).

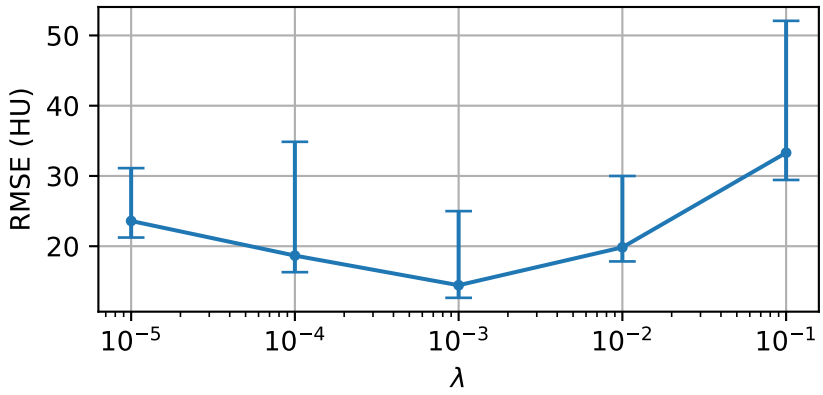
nificantly outperforms every other model pair-wise (p-value $< 0.5\%$).

Fig. 5.7.5a shows a region of interest centered around the needle in the axial slice with the highest errors of the first test subject with a misalignment of the prior scan by 2deg. FDKConvNet cannot recover fine structures and even inserts ghosting artifacts of the needle. The differences of Prior-Net and Prior-SegNet wrt. the ground truth are less pronounced. Prior-Net slightly blurs a small region around the needle whereas Prior-SegNet seems to insert a slight halo. Both are able to compensate for the misalignment of the prior.

Fig. 5.7.5b and c show the reconstruction errors of different models wrt. rotated prior scans. Prior-SegNet (no mis) performs significantly worse for $|\alpha| \leq 5.5$deg compared to the other models (see Fig. 5.7.5 c). Prior-SegNet performs best for $|\alpha| \leq 2.5$deg (see Fig. 5.7.5b) as well as $|\alpha| > 5.5$deg (see Fig. 5.7.5c), and Prior-Net for $2.5\text{deg} \leq |\alpha| \leq 5.5$deg (see Fig. 5.7.5b). Fig. 5.7.5 shows the reconstruction errors of the proposed network when choosing different values for λ_2 while keeping $\lambda_1 = 1$ fixed. The graph of RMSE median values and interquartile range in Fig. 5.7.5a clearly shows a minimum at $\lambda_2 = 10^{-3}$ which coincides with the empirically chosen value used in the previous experiments. This same minimum (or maximum in terms of SSIM and PSNR) can also be found in Fig. 5.7.5b showing the mean and standard deviation.

Comparing the median/interquartile range and mean/standard deviation for RMSE values gives further insights into the underlying distributions. Note that the median is always lower than the mean value, which indicates a skewed distribution towards lower

(a)



(b)

$\log_{10} \lambda_2$	SSIM [%]	PSNR [dB]	RMSE [HU]
-∞	95.71±3.31	41.09±3.50	28.70±15.09
-5	96.47±3.71	41.56±3.85	29.79±16.04
-4	95.11±7.39	40.95±6.67	39.44±48.51
-3	**97.15±3.47**	**43.97±4.95**	**23.21±18.11**
-2	96.31±4.48	41.97±4.66	29.73±21.88
-1	90.68±8.42	36.92±5.85	55.62±48.71

Figure 5.19: Reconstruction errors of Dual Branch Prior-SegNet wrt. segmentation loss scaler λ_2 ($\lambda_1 = 1$) on the test set (without misalignment). (a) RMSE vs. λ (median with interquartile range as error bars). (b) Mean and standard deviation of the different metrics.

values. The quartiles suggests the same: Q_1 is always closer to the median than Q_3. Also note that the standard deviation is always very high and in one case, i.e. $\lambda_2 = 10^{-2}$, even exceeds the mean value.

5.7.6 Discussion

Incorporating a patient-specific prior planning scan for interventional CBCT reconstruction from sparse projections – and therefore a drastically reduced amount of radiation that the patients and surgeons are exposed to – is a simple way to increase the quality of the reconstructions. FDKConvNet has no further information about the volume and can only correctly reconstruct (low frequencies of) tissues that are not significantly affected by streaking artifacts. Training with misaligned priors is essential to keep the quality at a high level. Though not trained with rotations $|\alpha| > 5$deg, Prior-Net and Prior-SegNet are able to compensate for up to 5.5deg. Confirming the initial hypothesis, the additional segmentation head of Prior-SegNet facilitates the reconstruction task for small angles of misalignment rotation and seems to generalize better for high angles.

Curiously, the outputs of the segmentation head for the chosen $\lambda_2 = 1\text{e}{-3}$ do not resemble actual segmentations of the interventional instruments at all and, consequently, cannot be used afterwards. This means that, although the segmentation loss has a positive influence on the quality of the reconstructions, the main focus of the optimization procedure is still on the reconstruction path. The additional trainings for finding an optimal λ_2 when keeping $\lambda_1 = 1$ fixed (see Fig. 5.7.5) revealed that the segmentation task is gradually taking over the reconstruction task for $\lambda_2 \geq 1\text{e}{-2}$. This can be seen in the loss values L_{MSE} and L_{Dice}: For $\lambda_2 < 1\text{e}{-2}$, L_{Dice} does not fall much below 1 and even slightly increases back towards 1 after the first 15 epochs. This relates to a Dice coefficient close to zero which, for obvious reasons, cannot result in a meaningful segmentation output. However, if $\lambda_2 \geq 1\text{e} - 2$, the segmentation loss gains enough importance in the network optimization such that L_{Dice} drops to 0.8 after 15 epochs and even 0.01 after 85 epochs. On one hand, this makes the output of the segmentation head sensible segmentations of the interventional instruments. On the other hand, this gained ability of the network simultaneously results in a degradation of the actual reconstruction output (see Fig. 5.7.5). Future use cases should therefore define whether the focus should be on reconstruction or segmentation task and set λ_2 accordingly. Since the aim of this work was to improve the reconstruction quality, $\lambda_2 = 1\text{e}{-3}$ is the optimal choice of the tested values for this data.

The observations of the lambda sweep in Fig. 5.7.5 show that the underlying distribution is probably not a normal distribution. For this reason, statistical tests that assume a normal distribution cannot be applied on this data. Therefore, Wilcoxon's signed-rank test was chosen in the previous significance tests instead of Student's t-test. The Mann-Whitney U test is another statistical test which does not assume a normal distribution but instead assuming independent observations from both groups. However, the ith sample of a group is generated by applying one of the methods to the ith input slice, which makes the sets of every ith sample dependent of each other. For this reason, the Mann-Whitney U test could not be used either.

Another important fact to note is that the values reported in the above tables and figures are the statistics of the entire test data set. This has two implications: (1) One cannot assume that the metrics are equal wrt. different subjects. In fact, it was found that the errors of few test subjects are considerably higher than the large part of all other subjects. These can be declared as outliers, but further tests have to show what causes them to be performing worse. (2) One cannot assume that the metrics are independent of the location of the (axial) slice. Reconstructions of cone beam projections from a circular trajectory suffer from decreasing quality with distance from the central slice. This is because Tuy's condition [Tuy83] is invalidated for circular trajectories, which results in unstable reconstructions. However, much information about the interventional scan can be taken from the prior acquisition. In fact, the errors do not increase in the outer axial regions but instead at the center. This can be explained by the position of the simulated needle which was inserted rather centrally and therefore makes it more difficult for the network to reconstruct these slices since it cannot simply take the data from the prior scan there. Moreover, the misalignment being an in-plane rotation simplifies the reconstruction task for the network in a way that it only needs to rotate the prior scan even if merely a small amount of information about the interventional scan is available which ultimately results in high quality reconstructions even in the outer axial regions.

In this work, the only type of prior scan misalignment was in-plane rotations. For application on clinical data, however, other types have to be considered and evaluated as well, e.g. translation, elastic distortions and breathing motion. Moreover, the interventional data in this work was simulated and without noise, which may have further effects when applying on real data.

The code is obtainable from Github[7].

[7] `https://github.com/phernst/prior-segnet`

Chapter 6

Failed Attempts and Error Analysis

In only a few cases, scientific research is based on random findings that solve a yet unknown problem or improve upon state-of-the-art results. In most cases, however, it is (supposed to be) based upon hypotheses built on prior knowledge as an attempt to solve a specific problem. These hypotheses must be evaluated scientifically and result in one out of two possible outcomes: acceptance or rejection.

In general, the main hypothesis of a novel method somehow corresponds to asking whether it is suitable to solve the problem to a certain degree (e.g. 'Does the method outperform the state of the art?' or 'Does the accuracy of the method exceed a certain threshold?'). If this main hypothesis has to be rejected, it is very unlikely for the method to be published since citations are biased towards publications of positive or statistically significant results which, in turn, biases the publications in journals and conferences [DUS+17; RB08].

Nevertheless, negative or statistically non-significant results advance research in a way that future research and resources do not need to be wasted in order to obtain the same (unpublished and therefore unknown to other researchers) results. Moreover, if novel methods had been developed, which these negative results were obtained from, they might be an interesting starting point or might contribute in brainstorming for the design of new methods in perhaps very different areas of research.

In this rather unconventional chapter, several methods that were developed in the course of this thesis and produced negative results are described briefly in the following sections for the aforementioned reasons. In addition, there is going to be an error analysis of each method attempting to identify possible reasons why the methods did not work as intended and the initial hypotheses had to be rejected.

6.1 Direct CT Reconstruction using CNN

6.1.1 Problem Statement and Hypothesis

The first idea that may come to mind when thinking about CT reconstruction with deep learning (in particular with CNNs) is feeding the network with the raw projection data directly from the scanner and predicting the reconstructed volume or image of attenuation values.

In fact, this has been tried successfully for MRI reconstruction using the AUTOMAP network [ZLC+18], which is a multilayer perceptron, essentially. In the scope of MRI reconstruction, the raw data is not projection values (like in CT reconstruction) but Fourier coefficients in the so-called k-space. Reconstructing the image data is usually performed by applying an inverse Fourier transform. Due to natural fluctuations and sampling patterns of the Fourier coefficients, different kinds of artifacts are contained in the reconstructed images if the inverse Fourier transform is applied directly. AUTOMAP attempts to learn a kind of improved inverse Fourier transform which automatically reduces/removes these artifacts from the images. Since it is a multilayer perceptron, the number of trainable parameters is in $O(M \cdot N)$ (M and N being the number of input and output nodes) – this is $O(n^4)$ for an $n \times n$ image (!). Although AUTOMAP achieves high quality reconstructions, it is obvious that it can only be applied on small image sizes (up to around $128\,\mathrm{px} \times 128\,\mathrm{px}$ already having $>\sim$260 million parameters).

As explained previously in Sec. 3.4, CT projections can be transformed to Fourier coefficients corresponding to a k-space from a radial MRI acquisition (in case of parallel beam CT). However, the scaling problem of AUTOMAP still remains, which makes it especially unusable for reconstructing relatively small objects, e.g. lymph node metastases or cancerous pulmonary nodules. CNNs, on the other hand, scale very well with increasing image sizes and have been used for state-of-the-art medical imaging methods before, as well. Therefore, it seems reasonable to use them for CT reconstruction.

6.1.2 Methods and First Technical Problems

However, trying to implement a CNN for this purpose already poses a first problem, even for the simplest case of two-dimensional image reconstruction from parallel beams: mismatching shapes between input, i.e. the sinogram, and output, i.e. the reconstructed image. Being based on discrete convolutions, CNNs are typically designed for equal input and output shapes (or up/downsampled versions by powers of 2). With an appropriate padding before or after applying the convolutions, it is possible to design a CNN which directly transforms sinogram data into image data, nevertheless (based on only the shapes of the data).

6.1.3 Results

Actually performing a training with such an architecture results in network predictions like the one depicted in Fig. 6.1. One can imagine this to be a good average of the reconstructions used for training. In fact, looking at the network predictions of different

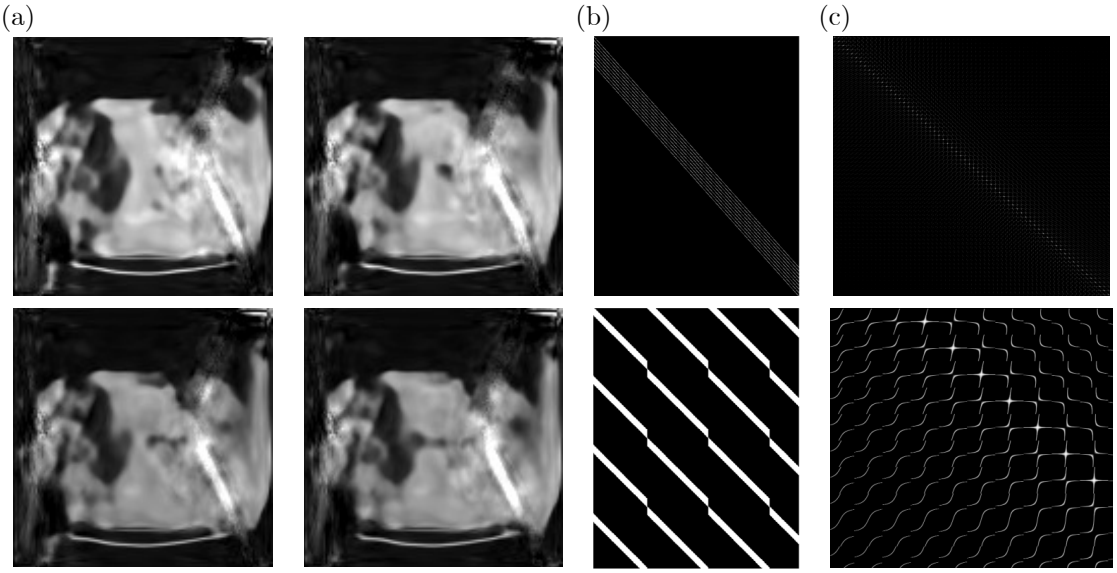


Figure 6.1: Direct CNN reconstruction. (a) Examplary outputs of the trained network. (b) Convolution matrix with a 9x9 kernel (top: full, bottom: zoomed). (c) System matrix of a parallel beam setup (top: full, bottom: zoomed).

sinograms, there is no significant visual difference between them, which means that the output is independent of the input.

6.1.4 Error Analysis

The reason for this failure is found in the mathematical nature of the convolutions. Originally designed to mimic the human vision, convolutional layers perform very local operations expressed by the kernel size. Therefore, the so-called receptive field is very limited. However, for CT reconstruction, a pixel in the reconstructed image depends not only on a local neighborhood of pixels in the sinogram but on pixels along a sinusoidal curve from all (ramp-filtered) projections.

This can also be observed when comparing the transposed CT system matrix (see Sec. 3.5), which corresponds to the backprojection matrix, and the convolution matrix, which is a doubly block circulant (Toeplitz) matrix. To make the backprojection (and therefore reconstruction) possible with a convolutional network, the elements of the convolution matrix need to cover at least the non-zero elements of the transposed CT system matrix:

$$\sum_i \sum_j H(|a_{ji} c_{ij}|) = \sum_i \sum_j H(|a_{ij}|) \tag{6.1}$$

using the Heaviside function $H : \mathbb{R} \to \{0, 1\}$, the CT system matrix $A \in \mathbb{R}^{M \times N}$ and the convolution matrix $C \in \mathbb{R}^{N \times M}$. To make Eq. 6.1 hold, the kernel sizes of the convolutions must be increased. However, since the per-pixel information is global wrt.

the sinogram, the necessary increase makes the CNN equal to a multilayer perceptron and the AUTOMAP architecture with the same scaling problems. This proves that CNNs cannot be (efficiently) used for direct CT reconstruction.

6.2 Primal-Dual Network and UNet with Fourier Transform Layers

6.2.1 Problem Statement and Hypotheses

The Primal-Dual Network and the architecture Primal-Dual UNet, explained in Sec. 5.2, were originally described for the image/reconstruction space as the primal space and the sinogram/projection space as the dual space using the Radon transform and the FBP as operations for switching between the spaces.

However, the primal-dual algorithm is not limited to these transformations, in general. Inspired by the Fourier slice theorem (see Sec. 3.4) and MRI reconstruction from frequencies in the k-space (see Sec. 6.1), the hypothesis is that using the image/reconstruction space as the primal space and the Fourier space as the dual space using the Fourier and inverse Fourier transform as operations for switching between the spaces works as well as or better than using Radon and FBP.

The idea is that (1) the (inverse) Fourier transform is commonly used in other imaging and wave processing contexts, and is therefore highly optimized in the form of (inverse) Fast Fourier Transform ((i)FFT) algorithms with minimal errors compared to the Radon transform and the FBP.

Additionally (2), CT reconstructions from undersampled projections (sparse views, low-resolution detectors due to detector binning, ...) usually lack frequency information (radial lines in case of sparse views or high frequencies in case of low-resolution detectors). For this reason, optimizing directly in Fourier space is supposed to have benefits for the reconstruction.

Finally (3), the shape of the reconstructed image (in the primal space) is equal to the shape in Fourier space (that is, the dual space) in contrast to using the sinogram/projection space as the dual space. This even suggests replacing the convolutional layers in the dual space with a UNet similar to the k-space learning network [HSY20].

6.2.2 Methods

Since the values in the Fourier space are complex (the frequencies consist of an amplitude and a phase encoded in the real and imaginary part of the complex number), there are different ways how to handle this in the CNN: (a) separating the real and imaginary part into two channels, followed by standard convolutional operations, then re-composing the complex number from the two channels (like in the k-space learning network [HSY20]), or (b) using the complex-valued counterparts of the convolutional operations (like in deep complex networks [TBZ+18]).

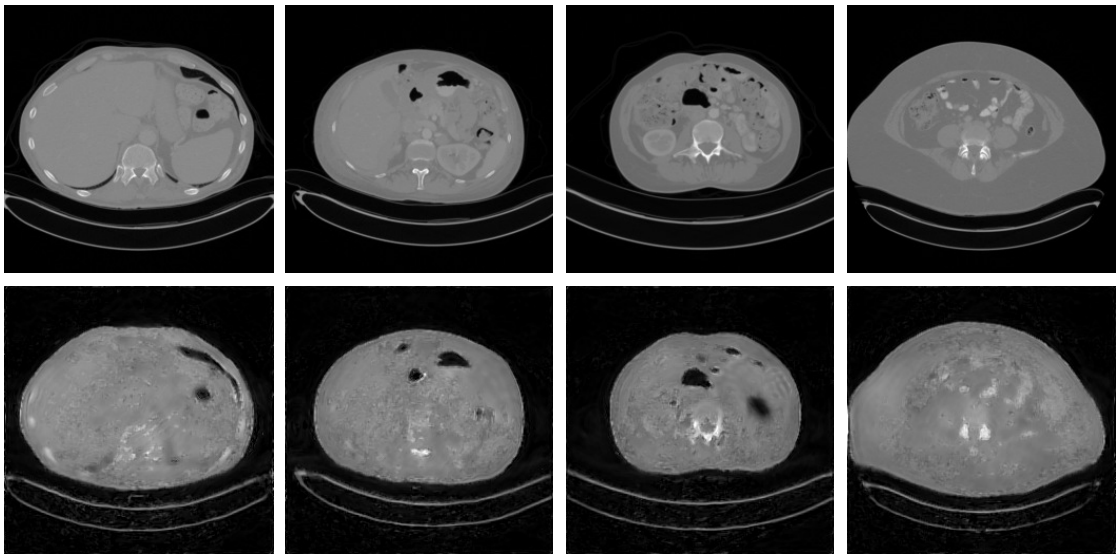


Figure 6.2: Four exemplary predictions of Primal-Dual UNet with Fourier dual space from 45 fan beam projections. Top: ground truth. Bottom: network prediction.

Moreover, one has to decide how the input to the network should be computed, i.e. how to transform the CT projections to the primal or dual (Fourier) space. As mentioned above, the Fourier slice theorem can help for this purpose. If parallel beam projections are available, a non-uniform Fourier transform algorithm (NUFFT) can be used to directly obtain the initial undersampled dual space. However, this kind of projections is very uncommon. Instead, fan (or cone) beam projections are typically acquired on CT scanners. The Fourier space of this type can be obtained from Fourier reconstruction methods [ZH95a]. Conversely, a primal space initialization can be computed using FBP or other common CT reconstruction algorithms which, however, seems counterintuitive when trying to model a reconstruction method based on only the Fourier space.

6.2.3 Results

Exemplary results of the Primal-Dual UNet with the Fourier dual space can be seen in Fig. 6.2. Interestingly, the network does not perform similar to the Radon-based Primal-Dual UNet, let alone outperforms it. Apart from the rather bad reconstruction quality in general, it even inserts a non-existing hole in the reconstruction in the third example. One reason might of course be the choice of hyperparameters of the network trainings but the following explanations are more likely.

6.2.4 Error Analysis

Starting with the Fourier initialization obtained from a Fourier reconstruction method, the explanation for failing is probably already found in the initialization. As pointed

out by the authors [ZH95a], the number of projections needs to exceed a certain limit dependent of the fan angle and the target resolution. Otherwise, ring artifacts are introduced in the Fourier space, getting higher in frequency the lower the number of projections. For the setup used in this experiment, the optimal number of projections to satisfy this condition is 925. However, only 22 projections were used (that is, less than 3 % of the necessary amount), which does not just fill the Fourier space with ring artifacts but likely does not even insert any useful information for the Primal-Dual UNet which gets this data as input. Therefore, using an FBP reconstruction as initialization in the primal space seems to be more sensible.

However, the suboptimal results of the networks initialized with the FBP reconstructions indicate that there is another problem which is in the network architecture.

Due to the radial structure of the projections in the Fourier space, the values are much more densely distributed around the zero-frequency and get sparser at higher frequencies, i.e. the Euclidean distance between projections in Fourier space increases with the frequency. Additionally, the amplitude of the frequencies decreases rapidly for most non-artificial images. These two properties invalidate the assumption of the locality of convolutions: the receptive field might not capture any value in the higher frequency areas, and the large differences in amplitudes depending on their position cannot be mapped in the convolution kernels.

Finally, a convolution in Fourier space equals a multiplication in image space and vice versa (*convolution theorem* [Bra86, pp. 108–112]). Since convolutional kernels are typically chosen to be very small, e.g. 3×3, a convolution in Fourier space with these kernels corresponds to a multiplication in image space with very low frequencies (that is, the inverse Fourier transform of the kernels). This does not seem to serve the purpose of optimizing the final reconstruction. However, it was shown in several publications [PRC+22; HSY20] that CNNs applied on Fourier space data are in fact able to reduce artifacts and improve the quality of MRI data, which might invalidate this concern.

The Pytorch implementation of the method can be downloaded from Github[1].

6.3 Cartesian Sinogram Upsampling using Delaunay Triangulation

6.3.1 Problem Statement and Hypothesis

The main problem of sparse view CT is the number of missing projections, usually invalidating the Nyquist theorem (as opposed to low-dose CT where the number of photons per detector pixel is too low to estimate the expected extinction value). For this reason, it is necessary to fill up this empty space with sensible values to reduce the dimensionality of the nullspace and therefore the ambiguity of the reconstructed images or volumes (see Sec. 3.5).

[1] `https://github.com/phernst/deep-fourier-fanbeam`

For helical – and in particular circular/axial – trajectories, the dimensions of the projection space can be separated into (fully sampled) detector dimensions and a (undersampled) trajectory parameter dimension, i.e. an angular dimension for circular/axial trajectories. Mathematically, the values of the detector pixels can be interpreted as a function of the angular dimension. Furthermore, if the domain of this function is interpreted as (a subset of) the real numbers where the actual projections are a discrete subset of this domain, then it is possible to construct new data points by finding and evaluating an *interpolant function*. Evaluation is a trivial task but finding the interpolant function is very challenging in this case of projection data. First choices usually include nearest-neighbor interpolation, linear interpolation, cubic spline interpolation or sinc interpolation (or their higher-dimensional counterparts). However, visual inspection of sinogram data already hints that these simple interpolants might not be the optimal choice:

1. Shapes in a sinogram can be described as sinusoidal (hence their name). However, these sinusoidal shapes are not the previously described functions but a kind of graph of them. This means that trying to express the functions as a single (shifted or scaled one-dimensional) sine function is not possible in general. Moreover, the simple interpolants are not able to preserve the sinusoidal shapes in the sinogram which already makes them bad interpolants for this reason.

2. Interpolating along the angular axis is one-dimensional, which is why only the one-dimensional variants of the simple interpolants can be used. However, since the projections are ray integrals over the attenuation coefficients, the values of other detector pixels, which are not necessarily in the local neighborhood, include information about the projection values to be interpolated.

From another point of view, the coordinates of values in the sinogram can be interpreted as polar coordinates: each value can be assigned an angle, i.e. the position along the angular dimension, and a distance from the origin, i.e. the position along one of the detector axes. This results in one more reason why there might be better choices than the simple interpolants:

3. Interpreting the positions of the sampled projections as polar coordinates results in similar properties pointed out in the explanation of the Fourier slice theorem for sparse views (see Sec. 3.4). Particularly, since each projection fills up the space by one radial line, the values around the origin are sampled much more densely than in the outer regions. This means that the quality of the interpolation decreases with distance from the origin if no further information about the values in this space is available. Optimally, the interpolant would also depend on the distance to the origin. This is obviously not taken care of by the simple interpolants.

Converting the (interpreted) polar representation of a sinogram to its Cartesian representation results in data similar to an (unfiltered) backprojection but without the "backsmearing", such that each point in this Cartesian form still represents exactly one

point of the polar form (as opposed to the backprojection where all points on a line contribute to one value in the sinogram).

These two properties – one-to-one point correspondences and backprojection-like images – build the hypothesis that interpolating missing sinogram values in the Cartesian representation is the optimal choice for sinogram interpolation.

6.3.2 Methods

The Cartesian representation makes it possible to use the two-dimensional (or multidimensional, in general) counterparts of the simple interpolants and therefore including more neighboring values for the interpolation of a point.

As experimentally determined by [rhtt], interpolating values on polar grids by using their Cartesian representation is most accurately done by using barycentric interpolation. For this purpose, a triangulation of the know data points, converted to Cartesian coordinates, is performed. In this case, the Delaunay triangulation algorithm [Del34] is used. To compute the interpolated value for a new data point P, the triangle containing P, represented by three points T_1, T_2 and T_3 of the triangulation, is identified first. For the purpose of this interpolation, the x- and y-coordinate of the points coincide with their respective Cartesian coordinates, whereas the z-coordinate is set to the data value. The new data point P is assumed to be in the plane of the three triangle points, i.e. P, T_1, T_2 and T_3 must be linearly dependent. Therefore, the following equation must hold [Lan86]:

$$\begin{vmatrix} P_x & P_y & P_z & 1 \\ T_{1x} & T_{1y} & T_{1z} & 1 \\ T_{2x} & T_{2y} & T_{2z} & 1 \\ T_{3x} & T_{3y} & T_{3z} & 1 \end{vmatrix} = 0$$

The only unknown variable is P_z, which corresponds to the interpolated value of the new data point. Solving for P_z results in the barycentric interpolation:

$$P_z = \frac{\begin{vmatrix} T_{1x} & T_{1y} & T_{1z} \\ T_{2x} & T_{2y} & T_{2z} \\ T_{3x} & T_{3y} & T_{3z} \end{vmatrix} - P_x \cdot \begin{vmatrix} T_{1y} & T_{1z} & 1 \\ T_{2y} & T_{2z} & 1 \\ T_{3y} & T_{3z} & 1 \end{vmatrix} + P_y \cdot \begin{vmatrix} T_{1x} & T_{1z} & 1 \\ T_{2x} & T_{2z} & 1 \\ T_{3x} & T_{3z} & 1 \end{vmatrix}}{\begin{vmatrix} T_{1x} & T_{1y} & 1 \\ T_{2x} & T_{2y} & 1 \\ T_{3x} & T_{3y} & 1 \end{vmatrix}}.$$

It is possible to optimize the interpolation further by replacing the Delaunay triangulation: Every new data point of the interpolation is surrounded by four known data points, spanning a trapezoid. The vertices of this trapezoid form four different triangles, and the new data point lies in at least two of them. This means that the values from different triangle interpolations can be combined and therefore result in possibly more accurate interpolations. However, this was not done for the following experiments.

Due to the radial nature of the polar coordinates in Cartesian space, further problems emerge that need to be handled properly:

- Interpolation of intermediate projections likely includes points outside the convex hull of the available sampled coordinates in the boundary region. These points are not inside any triangle of the Delaunay triangulation such that identifying T_1, T_2 and T_3 fails. To make these points part of the convex hull with as least additional computation as possible, the sparse sinogram is zero-padded along the detector axis such that the convex hull barely includes all points of the projections to be interpolated. In particular, the padding applied to both sides of the detector dimension can be determined as

$$P = \left\lceil \left(\frac{1}{\cos\left(\frac{\Delta\theta}{2}\right)} - 1 \right) \cdot \frac{D}{2} \right\rceil,$$

  where P is the padding on one side, D the original detector size and $\Delta\theta$ the angular difference between two consecutive projections.

- One main difference between a function of polar coordinates and a sinogram is their domains: by definition, the radius of a polar coordinate must be non-negative and the angular coordinate must be an element of an interval spanning 2π to be uniquely represented. These assumptions are not made for sinogram coordinates. Here, the detector axis is assumed to be centered about the origin (i.e. the rotation axis). Sinogram points with negative detector coordinates are converted to polar coordinates by using the symmetry property of the Radon transform: $\mathcal{R}f(s,\theta) = \mathcal{R}f(-s,\theta \pm \pi)$.

- For a unique representation of the points in a polar coordinate system, the pole is uniquely defined as well. This means that $|\{(0,\theta) : \theta \in [0,2\pi)\}| = 1$ which implies $f(0,\theta_1) = f(0,\theta_2)\,\forall\theta_1,\theta_2$ for a function f of polar coordinates. However, for a sinogram, it generally holds $\mathcal{R}f(0,\theta_1) \neq \mathcal{R}f(0,\theta_2)\,\forall\theta_1 \neq \theta_2$. To make all values of the sinogram available in its polar representation, the detector coordinates are shifted by half a detector pixel. This way, the pole is avoided for all sinogram points as well as the interpolated projections.

6.3.3 Results

Fig. 6.3 (left) shows the errors of the sinograms after upsampling using the different interpolation functions. The proposed Cartesian upsampling outperforms every other interpolation function in terms of RMSE if at least 6 projections are used. Linear upsampling is the second-best option in this case and is almost indistinguishable from Cartesian upsampling if more than 100 projections are used. Sinc interpolation results in slightly worse errors compared to linear interpolation. Finally, nearest neighbor upsampling is the least accurate method, which is expected because projections are continuous along the angular dimensions whereas nearest neighbor interpolation assumes step functions.

Rather surprisingly, the errors of the FBP reconstructions (with Hann window filtering) after upsampling are very different from the errors in sinogram/projection domain (see Fig. 6.3 (right)). Cartesian interpolation only slightly outperforms sinc interpolation

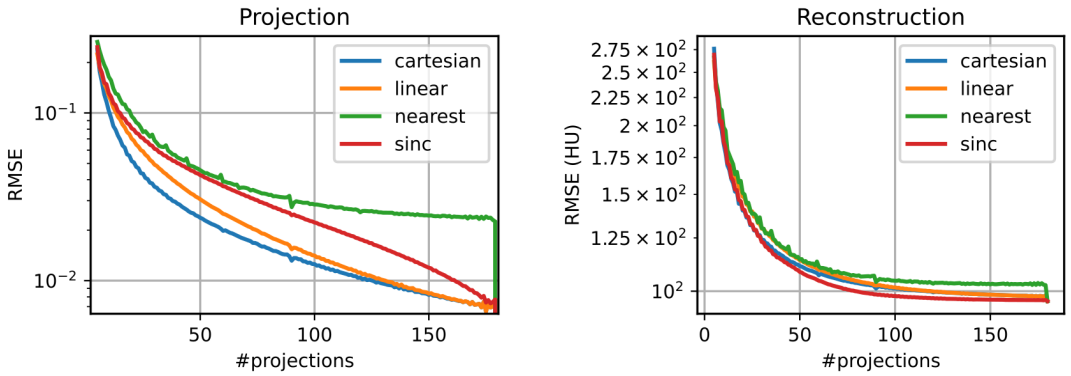


Figure 6.3: Errors of different fan beam sinogram interpolation functions in projection space (left) and after FBP reconstruction (right).

for up to 27 projections where sinc interpolation surpasses Cartesian interpolation and results in the lowest overall errors until the highest tested number of projections. Linear interpolation starts with errors similar to nearest neighbor interpolation but again approaches and becomes almost indistinguishable from Cartesian upsampling if more than 100 projections are used. Nearest neighbor interpolation still performs worst but is very similar to linear upsampling for up to 50 projections.

6.3.4 Error Analysis

Despite producing the lowest errors in sinogram domain, the FBP reconstructions of sinograms upsampled with the proposed Cartesian interpolation cannot outperform the reconstructions of sinc-interpolated sinograms. Since the reconstructions were created only using the FBP algorithm, it is not proven that other reconstruction methods perform the same using the interpolations of the Cartesian upsampling.

One should also note that the Cartesian sinogram interpolation was only described for parallel beam geometries. It is possible, however, to modify the algorithm to output sensible interpolations for fan beam geometries as well. This merely includes the conversion of the coordinates from the fan beam sinogram to a parallel beam sinogram [KS88, Sec. 3.4] before interpolation. Moreover, the interpolation of the negative and non-negative half-planes of the fan angle dimension should be carried out separately. This way, the Delaunay triangulation is more regular, comprising similar sizes of triangles (Fig. 6.4 (a) and (b)) as compared to computing the triangulation for the entire set of sampled fan beam sinogram points (Fig. 6.4 (c)), and thus results in more credible interpolations (compare Fig. 6.4 (e) and Fig. 6.4 (f)). However, the proposed upsampling method could not outperform the simple interpolants, in particular sinc-interpolation, for fan beam sinograms in neither projection nor reconstruction space.

Since the entire interpolation algorithm is algebraic, no rebinning or intermediate resampling is required. Nevertheless, care must be taken for the redundant values in the

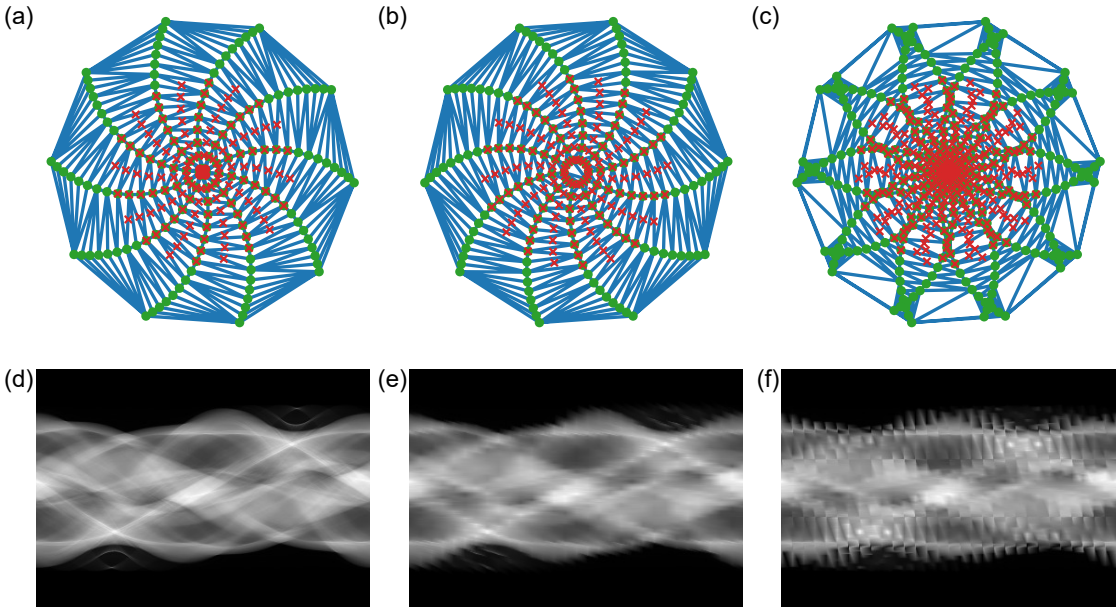


Figure 6.4: Cartesian sinogram interpolation for fan beam sinograms. Top row: sampled points after conversion to Cartesian space for $10 \rightarrow 20$ projections upsampling (green: sparse input sinogram (padded), blue: Delaunay triangulation, red: upsampled target sinogram). (a) Sampling of non-negative fan angle half-plane. (b) Sampling of negative fan angle half-plane. (c) Sampling of all fan angle coordinates. Bottom row: exemplary fan beam sinograms. (d) Ground truth. (e) 12x upsampling using separate half-planes. (f) 12x upsampling using full sampling plane.

fan beam sinogram if short scan trajectories are used, such that the interpolation stays stable.

The implementation of the Cartesian sinogram interpolation along with the evaluation wrt. the simple interpolants can be found on Github[2]. The repository includes interpolation methods for both parallel beam and fan beam samplings.

6.4 Cone Beam Projection Based Affine Registration

6.4.1 Problem Statement

Incorporating temporal or model prior knowledge for the reconstruction of interventional CT data usually requires some (at least coarse, see Sec. 5.6 and Sec. 5.7) kind of alignment between the interventional and the prior data: if planning scans are incorporated, they need to be registered to the interventional scan, if shape information about interventional instruments is incorporated, its parameters need to be estimated to resemble the correct location and alignment in the interventional scan, and so on.

There are many non-deep learning as well as deep learning based registration algorithms. The first kind usually relies on iterative optimization schemes to maximize the similarity between the fixed and the moving image/volume, given a parameterized transformation (e.g. rigid, affine or elastic). Despite being highly accurate in many cases, the main drawback of these registration algorithms is their runtime which, depending on the algorithm, data resolution and transformation type, can easily take several minutes up to hours. This makes them practically unusable during medical interventions and hinders their use as a pre-processing step for trainings of neural networks, especially if used in an online augmentation scenario.

On the other hand, several deep learning based methods have also been published. These have some immediately recognizable advantages over the traditional algorithms. Since many of them are based on CNNs, their processing time is much lower (often less than a second up to a few minutes), even if they imitate the traditional iterative algorithms by unrolling a few iterations. However, there are some disadvantages of this type of registration algorithms, too. Since they need to be trained on certain data, it cannot be assumed that they generalize properly which entails the necessity to evaluate these methods (with trained weights) on new data sets and possibly having to fine-tune. Contrarily, the traditional algorithms are not data-driven in that sense and can be expected to generate accurate alignments on all sorts of input data. Another downside of the deep learning methods is their accuracy, which is often only close to or worse than the traditional methods. This can be explained by the drastically reduced number of iterations and possibly the design choices of the network architectures.

However, the main requirement of a registration algorithm in an interventional setting is the processing time, since a slightly non-accurate alignment can be made up for by the subsequent reconstruction algorithm (see Sec. 5.6 and 5.7). For this reason, CNNs seem

[2] `https://github.com/phernst/cartesian-sinogram`

to be a good choice for this task and use case. They also enable end-to-end trainings, which result in lower loss values in many cases.

6.4.2 Methods and Hypothesis

There are several parameters that have to be chosen for a registration algorithm, depending on its use case. These do not only include the type of transformation or similarity metric but, especially in case of CT, what data to use for the optimization: the reconstructed volumes or the projections (or a combination). Many of the more general deep learning based registration algorithms perform the alignment based on the reconstructions. This seems sensible since the transformations are also performed in the reconstruction space.

In case of interventional CT, however, this becomes very challenging: due to the minimal dose that the subjects should be exposed to, the reconstructions contain severe artifacts, be it a very low signal-to-noise ratio from low-dose CT or very pronounced streaking artifacts from sparse view CT. Moreover, the attenuation values reconstructed from the interventional CBCT are not distributed like the values from the pre-interventional planning scan from a conventional CT, which poses an additional challenge for the choice of the similarity metric which therefore has to be able to handle multi-modal data properly.

What seems more promising in this case is a projection-based registration algorithm, i.e. the spatial transformations should be estimated from the projections, directly. This seems especially sensible for sparse view CT, since the projections themselves have a high quality and barely contain artifacts. The original projections of the pre-interventional scan, however, cannot directly be used (since their scanning geometry was different) and must be simulated. Furthermore, the projections lack one dimension which might make the registration task more difficult for the CNN to learn.

Depending on the network architecture, the neural network can contain a projection or backprojection layer, both of which are differentiable and are therefore able to correctly backpropagate the gradients during training.

The chosen architecture should meet the following requirements: one input contains the interventional projections (considered as fixed images), another input is the (reconstructed) pre-interventional planning scan (considered as moving volume). The output should be some parameterization of the spatial transformation. As the title of this section already suggests, the transformation to be considered is affine (in fact, rigid) but should be changeable to an elastic deformation field transformation without much effort.

These requirements make the following architectures conceivable:

1. Feature extraction in projection space: the interventional projections are treated as channels and are concatenated (along the channel dimension) with the simulated projections of the pre-interventional volume. This tensor of stacked projections is the input to a 2D CNN for feature extraction, followed by (a) a backprojection/reconstruction layer and a 3D CNN which regresses a displacement field, or (b) a regression CNN to regress the six parameters of an affine transformation.

2. Feature extraction in reconstruction space: the interventional projections are sent through a reconstruction layer and the resulting volume is concatenated (along the channel dimension) with the pre-interventional volume. This is the input to a 3D CNN for feature extraction, followed by (a) a displacement field or (b) a regression CNN to regress the six parameters of an affine transformation.

One should note that one implementation of the second approach with a displacement field as output (excluding the reconstruction layer at the beginning) is VoxelMorph [BZS+19], which performs as well as traditional registration methods for MRI brain registration, while being faster and more memory efficient, and can be considered a state-of-the-art model for medical image registration based on deep learning. Its implementation will be used as 'backend' for the proposed architectures.

Judging on the success of VoxelMorph, the hypothesis in this section is that one of the previous architectures is able to predict the parameters of the affine transform between a prior CT scan and interventional CBCT projections of the same subject to a degree that only a fine elastic registration would be necessary for an exact alignment, or in a range that a method like Dual Branch Prior-SegNet (Sec. 5.7) can handle.

6.4.3 Loss Functions

A very important part of the registration task is the choice of the loss function. Independent of the output of the network, a very simple choice is the similarity – or rather dissimilarity – (in terms of intensity values) between the fixed and the transformed moving image/volume.

In case of a displacement field as output, i.e. an *implicit* parameterization of the affine transform, this can be combined with a regularization of the displacement field, e.g. to favor or enforce smooth neighboring displacement vectors. The following experiments will use a weighted sum of MSE and gradient penalty loss, similar to VoxelMorph [BZS+19]:

$$L_a(f, m, \phi) = \underbrace{\frac{1}{|\Omega|} \sum_{\mathbf{p} \in \Omega} [f(\mathbf{p}) - [m \circ \phi](\mathbf{p})]^2}_{L_{sim}(f, m \circ \phi) = \text{MSE}(f, m \circ \phi)} + \underbrace{0.02}_{\lambda} \underbrace{\sum_{\mathbf{p} \in \Omega} \|\nabla \mathbf{u}(\mathbf{p})\|^2}_{L_{smooth}(\phi)},$$

with $\Omega \subset \mathbb{R}^3$ being the set of voxel positions, f being the fixed volume (here: the interventional reconstruction), m being the moving volume (here: the planning scan), $\mathbf{u}$ being the predicted displacement and $\phi = Id + \mathbf{u}$ like in [BZS+19]. This generally results in a type of unsupervised optimization since the transformation parameters can directly be inferred from the similarity between the fixed and the transformed moving volume. However, quantifying the registration errors in terms of the affine parameters is not straightforward when the only output is a displacement field. The following experiments will therefore employ a least squares approximation incorporating the singular value decomposition of the affine matrix from the (overdetermined system of equations of the) displacement field [AHB87] to be directly comparable to the models predicting the explicit parameterization.

In case of an *explicit* parameterization of the affine transform, the chosen loss function depends on the type of the parameterization: an affine matrix, Euler rotation angles + translation or quaternion + translation, to name a few possible choices. For the following experiments, the combination of quaternion + translation was chosen because it does not need any regularization and is smooth by definition [MAV17]. In contrast, Euler angles do not encode the rotation injectively and suffer from gimbal lock, and directly regressing the 3x4 matrix elements of an affine transformation does not result in a well-defined affine matrix, in general. This type of parameterization generally results in a type of supervised optimization since the similarity between the predicted and ground truth transformation parameters can directly be calculated. In this case, the loss function in the following experiments is going to be the MSE between the prior scan warped with the predicted and the ground truth affine transformation:

$$L_b(\mathbf{q}, \mathbf{t}, \hat{\mathbf{q}}, \hat{\mathbf{t}}) = \frac{1}{|\Omega|} \sum_{\mathbf{p} \in \Omega} [m(\mathbf{q}\mathbf{p}\mathbf{q}^{-1} + \mathbf{t}) - m(\hat{\mathbf{q}}\mathbf{p}\hat{\mathbf{q}}^{-1} + \hat{\mathbf{t}})]^2,$$

with $\mathbf{q}$ and $\mathbf{t}$ being the predicted quaternion and translation, and $\hat{\mathbf{q}}$ and $\hat{\mathbf{t}}$ being the respective ground truth variables. Note that the moving volume warped with the ground truth transformation could also be replaced by the fixed volume, but this introduces unwanted errors (e.g. if the intensity distributions are different or the fixed volume contains artifacts). The loss function does not directly calculate dissimilarity between the parameters of the affine transform but instead indirectly from the volumes warped with these parameters. This way, the optimization is only based on image data and the networks learn by themselves how to set the quaternions and translation vectors properly.

6.4.4 Data Sets and Preprocessing

For the experiments, the Mayo Clinic data set was used because it produces the least amount of truncation artifacts and seems to be a good choice for affine alignment, since the skull is usually not deformed elastically (as opposed to the abdominal area in the CT Lymph Nodes or LungCT-Diagnosis data sets). To save memory, the volumes were downscaled to $128 \times 128 \times 128$ voxels with a voxel size of 1.5 mm and 13 equi-angular X-ray projections were simulated for each volume along a circular trajectory with a detector size of 512 px $\times$ 512 px with detector elements of 0.616 mm $\times$ 0.616 mm. During training, random flips of the sagittal axis were performed as augmentation. Each network was trained with misalignments of up to $\pm 10°$ and ± 38.4 mm and combinations thereof for each axis. Adam was chosen as the optimizer with a cosine annealing learning rate scheduler starting at 1×10^{-4} decaying to 0 after 1000 epochs with an effective batch size of 16. Early stopping on the validation loss was performed to choose the optimal model.

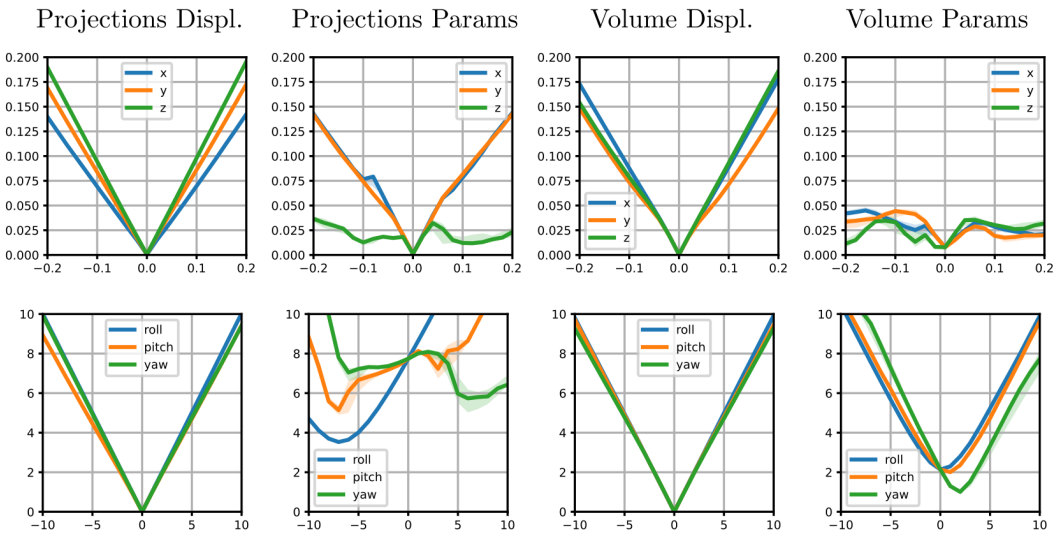


Figure 6.5: Affine registration errors. Top row: translation errors. Bottom row: rotation errors as geodesic distances.

6.4.5 Results

The registration errors for all four described combinations can be found in Fig. 6.5. Generally, none of the methods is able to compensate rotation sufficiently. While the methods outputting a displacement field slightly decrease the respective geodesic distances or do not change them, the methods outputting the affine parameters increase these distances in most cases and do not even result in a zero distance when no rotation is applied to the prior scan at all. The translation is handled better. The methods outputting a displacement field are able to reduce the translation error slightly better than the geodesic distances for rotations. Architecture 1(b) extracting the features in projection space and outputting the affine parameters compensates well for translations along the cranial-caudal axis (3.87 mm on average) but worse for the other axes. The method that compensates best for translation errors is Architecture 2(b) extracting the features in reconstruction space and outputting the affine parameters: the translation errors for every axis are all below 11.27 mm, and 5.23 mm on average. This renders Architecture 2(b) the optimal choice for this data under these training circumstances, at least wrt. translation.

Since none of the methods is able to compensate for rotation errors sufficiently, the initial hypothesis gets invalidated. Despite Architecture 2(b) being able to compensate for translation errors well, it introduces rotation errors of at least 1° in every case.

6.4.6 Error Analysis

Unfortunately, a direct comparison of these errors with the original VoxelMorph is not possible because the authors decided to assess the errors by comparing segmentation masks after registration in terms of the Dice coefficient. Those kinds of masks, however, were not available for the proposed CT registration. Due to the large errors in Fig. 6.5, one can assume the errors of the proposed architectures to be much worse than those that could be achieved in VoxelMorph. There are several possible explanations for this failure.

The first obvious difference is the data set. Where VoxelMorph acts on MRI scans of brains, the proposed architectures are trained on CT data of entire heads. For this reason, the intensity distributions are fundamentally different (between CT and MRI) and the networks need to be able to align tissues with very high attenuation values, like cranial bones, as well as tissues with rather low attenuation, like gray and white matter. The architectures outputting displacement fields failed this distinction by mainly considering highly absorbing tissues (see Fig. 6.6), such that the estimation of the affine parameters inevitably failed, too. Another difference in the data set, which might have a significant impact on the registration quality, is its size. Where VoxelMorph uses 3231, 250 and 250 different scans for training, validation and test set, the proposed architectures only use 35, 8 and 7 subjects, respectively (though including augmentation in contrast to VoxelMorph).

A further difference is the type of the registration transformation. VoxelMorph is trained to perform deformable registration (assuming an already performed affine alignment in advance), whereas the proposed methods are trained for affine registration. One could hypothesize that CNNs are able to handle (local) deformations better than (global) affine transformations due to the receptive field of the convolutional layers. Moreover, expressing the transformation in terms of a displacement field might be suboptimal when attempting to predict affine transforms: rotations modify the displacement vectors depending on their position, requiring global knowledge of the volume, in contrast to translations which are equally distributed to all displacement vectors. This might also explain why the rotation errors are high for every tested architecture.

Finally, major errors are likely to be introduced by the sparse views of the interventional CT. When extracting the features in projection space, both sets of interventional and simulated prior scan projections contain the same low number of views such that corresponding projections can be matched and processed by the convolutional layers. However, this implies that much information of the prior scan remains unused, i.e. the projection data that could have been simulated from positions in between the available interventional projections. On the other hand, when extracting the features in reconstruction space, the sparse interventional projections need to be reconstructed first, which necessarily introduces severe streaking artifacts, such that it becomes significantly more difficult for the network to extract the correct information.

These explanations show that the registration of sparse view interventional CT scans with CNNs is not a straightforward nor a trivial task. Despite building upon the state-of-the-art VoxelMorph architecture, the results of the proposed methods are less than

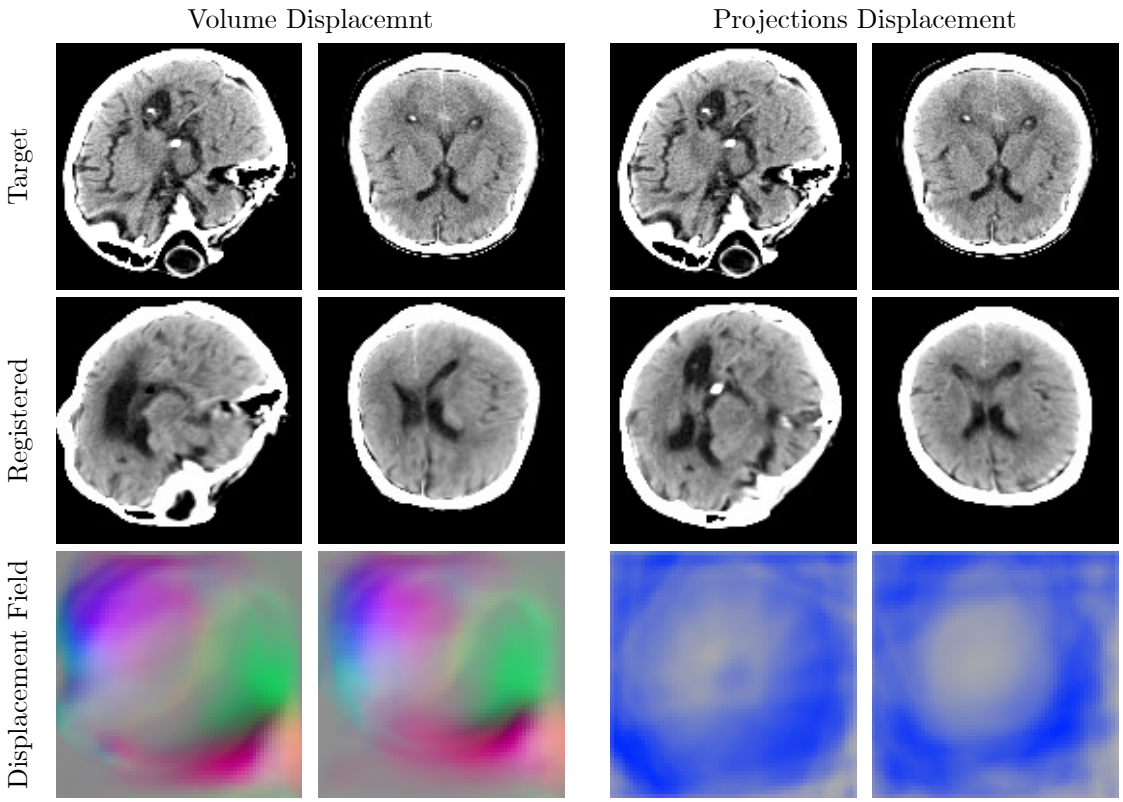


Figure 6.6: Two exemplary results for the methods using displacement fields for registration.

satisfactory. However, the field of deep learning based image registration is large and still growing, iterative approaches have not been evaluated in this thesis and traditional algorithms might also suffice as a preprocessing step. Since the focus of this thesis is on reconstruction, a thorough investigation of sparse view CT registration algorithms is waived here.

Chapter 7

Conclusion

This final chapter concludes the thesis. To this end, the first chapters will be summarized briefly before the methods presented in Ch. 5 will be compared among each other and discussed wrt. the goals of the thesis. Moreover, the research questions stated in Ch. 1 will be answered explicitly in the following section. Finally, the current limitations of the presented methods will be described and a short overview of ongoing and future work will be given.

7.1 Summary and Discussion

The main objective of this thesis is to reduce the X-ray exposure of surgeons and patients during CT-guided interventions by means of acquiring fewer projections while retaining a high quality of the reconstructions with the help of prior knowledge and deep learning, CNNs in particular. After introducing and motivating this task in the first chapter, the historical developments of X-ray tomography and CT imaging as well as deep learning and CNNs were described to get an impression of the technical and device-specific capabilities and limitations for this task. The mathematical background of CT image reconstruction and related concepts was given in the subsequent chapter to gain a deep understanding of how the algorithms work, followed by the state-of-the-art methods separated by the different types of prior knowledge. Each of the methods presented in Ch.5 is an example for incorporating prior knowledge of (mainly) a certain type for the task of sparse view CT reconstruction using CNNs. In that chapter, however, the individual methods are merely compared to and discussed wrt. other state-of-the-art algorithms and not among each other or in the scope of prior knowledge. Nevertheless, this is necessary to answer the first research question. Recall (and see Sec. 1.3 for detailed descriptions and explanations):

Research Question 1 *How do the three different types of prior knowledge – Algebraic, Deep Learning and Temporal/Model Prior Knowledge – influence the quality of the final reconstructions?*

The direct comparison of the presented methods is not straightforward. This is, among

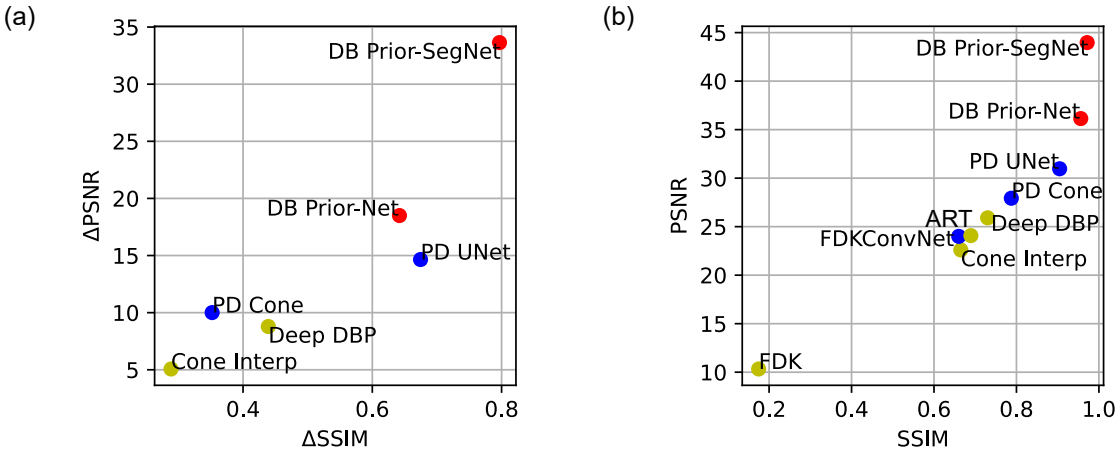


Figure 7.1: PSNR and SSIM values of different methods/models. (a) Increase wrt. FBP/FDK reconstructions. (b) Absolute values. Colors encode the primal type of prior knowledge: Algebraic (yellow), Deep Learning (blue) and Temporal and Model Prior Knowledge (red).

other reasons, due to the type of prior knowledge which is incorporated, e.g. methods incorporating planning scans were evaluated on interventional scans including needles whereas other methods were tested without any additional instruments, the dimensions of the processed data, i.e. stacks of 1D or 2D projections, and reconstructed 2D slices or 3D volumes, and the sparsity of the projections, i.e. some methods only allow subsampling by powers of 2 while others can be sampled arbitrarily. Despite these differences, it was attempted to keep these parameters as similar as possible for the sake of comparability. For this reason, one can get at least a rough impression of the effects of the types of prior knowledge on CT image reconstruction. Visual inspection of the resulting exemplary reconstructions of the individual methods in Ch. 5 already hints that Temporal/Model Prior Knowledge seems to be the winner among the investigated types. To support this assumption quantitatively, the SSIM and PSNR values of the different methods are shown in Fig. 7.1, where larger values indicate higher similarity. For the reasons described above, it is not meaningful to report specific values here but rather to describe trends.

In Fig. 7.1(a), where the increase of the metrics wrt. to their corresponding FBP/FDK reconstructions is plotted, one can easily identify Temporal/Model Prior Knowledge in the top right corner, Deep Learning Prior Knowledge situated rather centrally and Algebraic Prior Knowledge in the lower left corner. This reflects the results of the previously described visual inspection: Temporal/Model Prior Knowledge in fact results in the highest quality gain while Algebraic Prior Knowledge improves the reconstructions the least (among the three assessed types with the described methods, not considering other potential types). Note, however, that none of plotted points has negative values, which indicates that all methods do improve the FBP/FDK reconstruction, which serves as the common ground for comparison here.

Fig. 7.1(b) shows the absolute PSNR and SSIM values of the methods and also includes the post-processing CNN FDKConvNet as well as the algebraic iterative algorithm ART, both of which having a comparatively similar quality to the cone beam projection interpolation network presented in Sec. 5.4. Still, these methods are at the lower end of quality improvement compared to the other proposed methods in this thesis.

One might also note that the order of the methods in Fig. 7.1(b) is slightly different to the order in Fig. 7.1(a). This is because the values in Fig. 7.1(a) were calculated wrt. the FBP/FDK results of the individual methods (which, again, differ due to the previously described parameters as well as the data sets they were applied on) whereas Fig. 7.1(b) simply reports the absolute values, which underlines that only general trends should be considered instead of actual values.

Research Question 2 *How well do different similarity metrics assess the quality of reconstructed images/volumes wrt. a specific task, and which metrics are most suitable for evaluating CT reconstruction quality?*
Sec. 3.8 introduced and defined the most commonly used error and similarity metrics for CT image reconstruction, most notably MSE, PSNR, SSIM and VGG. One common problem of these metrics is their ranges. Despite CT images having defined physically interpretable scales, i.e. Hounsfield units or (mass) attenuation coefficients which relate image values to physical properties and in the best case directly identify a certain type of tissue, none of these metrics reports errors in one of these scales: MSE squares the values, PSNR's scale is in decibels (wrt. a target image), SSIM is usually between zero and one, and VGG's output is non-negative without a specific (or directly interpretable) range. Only MAE retains the scale of the images and therefore enables drawing conclusions about errors directly, however it is reported less frequently in publications than e.g. MSE.

Modifying the MSE slightly, though, to become the RMSE "resets" the resulting scale back to the original one, and is therefore favored over the simple MSE (due to the gained interpretability). Another way to make the MSE somewhat interpretable is normalization, resulting in NMSE or NRMSE. These metrics are usually specified in percents and can be interpreted as a kind of average percentual error of an image wrt. a target image. The main disadvantage of all of these MSE-related metrics is their proneness to outliers: since the mean value is calculated from all squared errors, one single outlier can drastically influence the final value (especially due to being squared). Depending on the use case, this can be wanted or unwanted behavior: For interventional scans including a needle, the MSE would be high if the reconstruction method removes the needle from the image, which is wanted behavior. If, on the other hand, the reconstruction of a conventional CT scan contains Poisson noise such that some pixel values are significantly hyper- or hypointense, the MSE would be high as well, although most areas are depicted sufficiently close to the ground truth image, which is unwanted behavior. An example is shown in Fig. 7.2. Further information about how scaling image values influences the MSE can be found in Apx. B.1.

The PSNR is derived from and therefore closely related to the MSE/RMSE. This

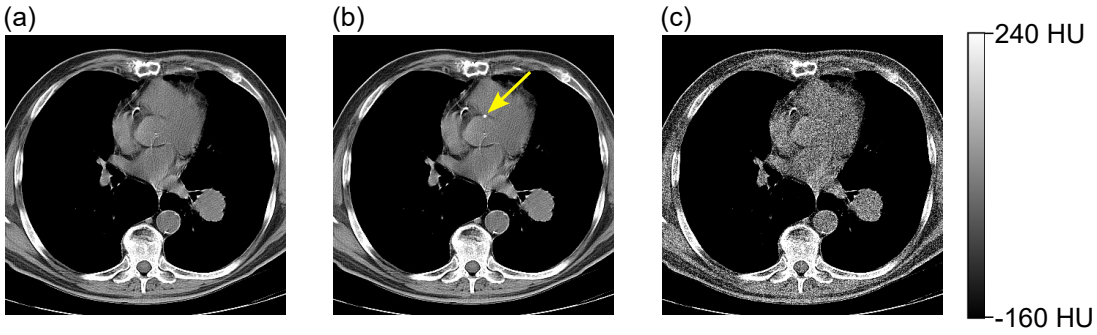


Figure 7.2: Example for effects of interventional needles and Poisson noise on MSE. (a) Conventional CT scan. (b) Like (a) but including an interventional needle (indicated by the arrow). (c) Like (a) but with Poisson noise. The RMSE between (a) and (b) is the same as the RMSE between (a) and (c), i.e. 80 HU.

makes it inherit the same disadvantages of MSE and adds the disadvantage of the scale being in decibels. Moreover, the calculation of PSNR includes defining a maximum peak intensity value. In case of CT images, such a value is not uniquely defined and therefore needs to be set sensibly. This, however, hinders comparability of different publications due to likely differently set values. It can be shown (see Apx. B.2), though, that this value merely biases the PSNR additively such that a PSNR based on a different maximum peak intensity value can easily be corrected in retrospect. Furthermore, ordering and differences unchanged.

SSIM has the advantage of incorporating luminance, contrast and structure in its calculation and is therefore a metric evaluating images more similar to human perception. In many cases, this results in quantitative ratings correlated to subjective human raters [LB09]. This makes it especially suitable for evaluating the quality of a reconstructed image. However, one downside the range between zero and one, which is not directly traceable back to the intensity values of the reconstructions. Moreover, it is necessary to set a dynamic range value. Setting it too large pushes the values towards one while setting it too low might result in unstable calculations or even zero division. As for PSNR, finding a sensible value for this parameter is not straightforward for CT images and hinders comparison with other published methods, as well. Even worse, there is no simple mathematical relation between SSIMs calculated with differing dynamic range values and a retrospective correction is not possible due to the SSIM not being injective wrt. this parameter (shown in Apx. B.3).

Though only briefly described and not further used in the thesis, VGG is likely to be metric most closely related to human perception since it is calculated from features that are able to classify images. The quantitative values, however, do not have a sensible meaning and also cannot directly be traced back to the image intensities. Since it uses weights of a VGG network trained on natural images, it is also questionable how well the values translate to CT images.

Finally, having humans evaluate the reconstructions is – scientifically – the best way

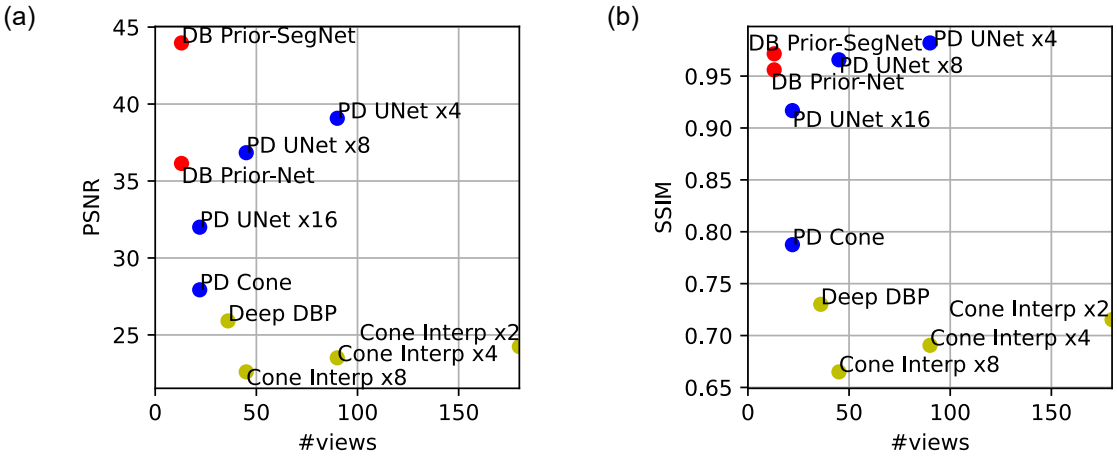


Figure 7.3: Metrics of different methods/models depending on view count. (a) PSNR. (b) SSIM. Colors encode the primal type of prior knowledge: Algebraic (yellow), Deep Learning (blue) and Temporal and Model Prior Knowledge (red).

to assess the quality. Not only does this produce the most meaningful values, but it can also be tailored to a specific task. The obvious downsides, however, are the scarce availability of doctors for this very time-consuming task (especially for large data sets), the necessity to clearly define the evaluation process to avoid large inter-rater variability making the results less or unusable, and the design of the tool that is used for the evaluation which must avoid ambiguities.

Research Question 3 *How does the reduced X-ray exposure (by reducing the number of projections) in combination with the incorporated prior knowledge correlate with reconstruction quality and computation time, and what does this mean for medical applications?*

Fixing a certain reconstruction algorithm while varying the number of available projections naturally results in lower quality for fewer projections and higher quality for more projections, which is an obvious observation since fewer projections means less available information (and therefore a higher-dimensional nullspace). In Fig. 7.3, this is directly apparent for the different upsampling factors of *PD UNet* and *Cone Interp*: the lower the upsampling factor, i.e. the more projections were available, the higher the PSNR and the SSIM.

On the other hand, fixing a certain number of projections while varying the different reconstruction algorithms shows no correlation wrt. the similarity metrics. With the values plotted in Fig. 7.3, one might even assume an inverse correlation: *DB Prior-SegNet* has the highest PSNR value with only 13 projections, while *Cone Interp x2* has one of the lowest PSNR values with 180 projections. This, however, is a false assumption because some presented methods have not been evaluated with a higher projection count (as this would simply increase the quality, as described in the previous paragraph) and

Table 7.1: Composition of the different methods. Arrows describe computation time: (↓↓) very low, (↓) low, (↑) high, (↑↑) very high. Colors encode the primal type of prior knowledge: Algebraic (yellow), Deep Learning (blue) and Temporal and Model Prior Knowledge (red).

Method	Composition	Time
● Sparse FDK/FBP	Sparse FDK/FBP (↓↓)	(↓↓)
● Full FDK/FBP	Full FDK/FBP (↓)	(↓)
● ART	(Backprojection, Update, Projection)↻	(↑/↑↑)
● Cone Interp (x8)	3x CNN (↓), Full FDK (↓)	(↓)
● Deep DBP	Hilbert (↑↑), Sparse FDK (↓↓), CNN (↓), Blend (↓)	(↑↑)
● PD UNet	4x CNN (↓), 3x Sparse FBP (↓↓), Projection (↓↓)	(↓)
● PD Cone	4x 3D CNN (↑), 3x Sparse FDK (↓↓), Projection (↓↓)	(↑)
● DB Prior-Net	Sparse FDK (↓↓), CNN (↓)	(↓)
● DB Prior-SegNet	Sparse FDK (↓↓), CNN (↓)	(↓)

are therefore not plotted in Fig. 7.3.

The last parameter, which is especially important for interventional applications, is the total computation time for the final reconstruction. Due to the different imaging parameters of the presented methods and varying use of reconstruction libraries, this, again, is not directly comparable. However, analyzing the composition of the algorithms, seen in Tab. 7.1, already reveals the ranking in terms of reconstruction time. Since the computation time is very low for Sparse FBP/FDK and low for (2D) CNNs and Full FBPFDK, the methods that only use these components are the fastest. These are *Cone Interp*, *PD UNet*, *DB Prior-Net* and *DB Prior-SegNet*. 3D CNNs are slower due to the additional dimension that they have to process. Therefore, *PD Cone* is slower than the previously mentioned methods. Finally, *Deep DBP* is very slow, where the computation of the Hilbert planes is the bottleneck. For comparison, *ART* was also included. Due to its iterative nature, it is usually slow or very slow, depending on the number of iterations and the update step.

Combining the three parameters *image quality*, *projection count* and *reconstruction time*, the type of prior knowledge that leads all of them is *Temporal and Model Prior Knowledge*: with only 13 projections and a low computation time, *DB Prior-SegNet* is able to achieve the best average PSNR of 43.97 dB and SSIM of 97.15 % on the LungCT-Diagnosis data set, which renders this type of prior knowledge most suitable for an actual application in a medical interventional setting. This does not imply that the other types are not well suited. Future methods that include all three of the described types of prior knowledge to an equal amount might mutually benefit from the additional information and, if well-designed, optimize the reconstructions in terms of the three parameters that were discussed here. This, combined with further optimizations like allowing more flexibility and variability of the data that the methods process, is bound to improve the work flow during medical interventions.

7.2 Current Limitations

The methods presented in this thesis are merely a few further steps in the direction of achieving CT-guided surgery with a minimum amount of X-ray exposure for both the patients and the surgeons. As already pointed out in Sec. 1.4, the main focus of the thesis is methodological developments for incorporating different types of prior knowledge into CT reconstruction algorithms, where clinical practical application is subordinate. Moreover, limiting the types of prior knowledge to Algebraic, Deep Learning and Temporal/Model Prior Knowledge only evaluates a subset of all possible types. This is even reinforced by a lack of systematic reviews of these types in the scientific literature for this field (e.g. defining all types and/or hierarchical relations between them, as done for the components in Fig. 1.1). Furthermore, the presented methods for the investigated prior knowledge types are by far not exhaustive. It is, e.g., conceivable that incorporating different Algebraic Prior Knowledge might improve the reconstructions significantly or incorporating different Temporal/Model Prior Knowledge might result in a quality drop compared to what is currently achieved.

In addition to the limitations pointed out in Ch. 1, which set the frame for the thesis as a whole, the individual methods defined further limitations for the time being. Due to a lack of large CBCT data sets (let alone interventional scans), all cone beam projections, which were used for training the networks, have been simulated. This, of course, does not take into account many types of artifacts. On the one hand, this makes the trained networks not directly applicable to real data without fine-tuning. On the other hand, these idealized settings allow for a more sensible evaluation of the actual modifications in the methods compared to the baselines since they are not influenced by artifacts in the data. Examples for these idealized settings include untruncated projections (as in Sec. 5.4 where, unrealistically, the entire abdomen is depicted on the detector), noiseless projections (i.e. values of the detector pixels are the expected values instead of samples of random variables depending on the energy of the photons from the X-ray tube), or perfectly aligned planning and interventional scans (which is impossible in real scenarios because of, e.g., breathing motion).

Nevertheless, some types of artifacts were introduced which might be less pronounced or absent in real data, to make the network training less time- and/or memory-consuming or because of requirements of the algebraic reconstruction algorithms. These include a low resolution of the detector (in most cases subsampled by a factor of at least 2) or the circular scan trajectory (providing insufficient data to the FDK algorithm and therefore resulting in cone beam artifacts).

7.3 Future Work

Summarizing the previous section, it becomes evident what future work has to focus on. Many of the limitations can be loosened without much effort by simulating the projections more realistically, e.g. including photon noise and increasing the detector resolution to match the size of real detectors, however significantly increasing time and

memory requirements, or by increased availability of (interventional) CBCT data sets (possibly including the raw projections) for training and therefore waiving simulations entirely (or merely using them as a means of data augmentation).

Furthermore, despite rejecting the respective initial hypotheses, the methods described in Ch. 6 can be reference points for future algorithms and show which errors might be avoided when designing them.

Since the ultimate goal of this research is clinical application, it is not only necessary to make the algorithms work on real data but also have them evaluated quantitatively by doctors and clinicians to investigate how beneficial the improvements of the methods in fact are during surgery. Moreover, it must be evaluated how the methods are supposed to be included in the workflow of the surgeons, which likely depends on the type of surgery that is done and therefore might need to be adjusted for certain use cases.

From a broader perspective, improving CT-guided interventions is not bound to incorporating data from CT only, despite being one of the quickest yet highest quality imaging modalities. In fact, surgeons mainly need to be able to navigate their instruments to the correct places inside the body while being able to distinguish between different types of tissues and materials. Although CT is a very good choice for this purpose (not considering the harmful radiation), other modalities can be supportive for this task as well, e.g. sonography or MRI. Acquisitions from these different imaging modalities could be combined to not only reduce the X-ray dose but also to gain additional information that would otherwise be unavailable when only relying on CT data, like information about blood flow from Doppler sonography without additional contrast agents or enhanced contrast of certain tissues with specific sequences of MRI. Software for robot-assisted surgery might also benefit from the combined imaging information to automate minimally invasive procedures more reliably and therefore reducing the risk of human errors. Moreover, recent advances in ultrasound imaging include Ultrasound CT (USCT) as a risk-free alternative to X-ray CT with high soft tissue contrast but different limitations. Due to their mathematical similarity, the methods proposed in this thesis might also be used for USCT reconstruction, eventually, entering an entirely new world of applications.

Bibliography

[AHB87] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-d point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(5):698–700, 1987. DOI: 10.1109/TPAMI.1987.4767965.

[AK84] A.H. Andersen and A.C. Kak. Simultaneous algebraic reconstruction technique (sart): a superior implementation of the art algorithm. *Ultrasonic Imaging*, 6(1):81–94, 1984. ISSN: 0161-7346. DOI: 10.1016/0161-7346(84)90008-7. URL: https://www.sciencedirect.com/science/article/pii/0161734684900087.

[AÖ18] Jonas Adler and Ozan Öktem. Learned primal-dual reconstruction. *IEEE Transactions on Medical Imaging*, 37(6):1322–1332, 2018. DOI: 10.1109/TMI.2018.2799231.

[BB11] Jerrold T Bushberg and John M Boone. *The essential physics of medical imaging*. Lippincott Williams & Wilkins, 2011.

[BBF20] Ivana Blažić, Boris Brkljačić, and Guy Frija. The use of imaging in COVID-19—results of a global survey by the international society of radiology. *European Radiology*, 31(3):1185–1193, September 2020. DOI: 10.1007/s00330-020-07252-3.

[BBP+19] Dimitrios Bellos, Mark Basham, Tony Pridmore, and Andrew P. French. A convolutional neural network for fast upsampling of undersampled tomograms in X-ray CT time-series using a representative highly sampled tomogram. *Journal of Synchrotron Radiation*, 26(3):839–853, May 2019. DOI: 10.1107/S1600577519003448.

[Beh15] Rolf Behling. *Modern Diagnostic X-Ray Sources: Technology, Manufacturing, Reliability*. CRC Press, first edition, 2015. DOI: 10.1201/b18655.

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer New York, NY, first edition, 2006. ISBN: 978-1-4939-3843-8.

[Bra86] Ronald Newbold Bracewell. *The Fourier transform and its applications*. McGraw-hill New York, 1986.

[BS93] C. Bouman and K. Sauer. A generalized gaussian image model for edge-preserving map estimation. *IEEE Transactions on Image Processing*, 2(3):296–310, 1993. DOI: 10.1109/83.236536.

[Buz11] Thorsten M. Buzug. *Computertomographie (ct)*. In *Medizintechnik: Verfahren – Systeme – Informationsverarbeitung*. Rüdiger Kramme, editor. Springer Berlin Heidelberg, 2011, pages 317–337. ISBN: 978-3-642-16187-2. DOI: 10.1007/978-3-642-16187-2_18.

[BZS+19] Guha Balakrishnan, Amy Zhao, Mert Sabuncu, John Guttag, and Adrian V. Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE TMI: Transactions on Medical Imaging*, 38:1788–1800, 8, 2019.

[Cam] Cambridge Dictionary. Knowledge. In *The Cambridge Dictionary Online*. Cambridge University Press. URL: https://dictionary.cambridge.org/us/dictionary/english/knowledge (visited on 02/25/2022).

[CDR99] Harrell G. Chotas, James T. Dobbins, and Carl E. Ravin. Principles of digital radiography with large-area, electronically readable detectors: a review of the basics. *Radiology*, 210(3):595–599, 1999. DOI: 10.1148/radiology.210.3.r99mr15595. PMID: 10207454.

[Cer16] Guillermo Avendaño Cervantes. The basics of x-rays. In *Technical Fundamentals of Radiology and CT*, 2053-2563, 1-1 to 1–5. IOP Publishing, 2016. ISBN: 978-0-7503-1212-7. DOI: 10.1088/978-0-7503-1212-7ch1.

[CK08] Marjeta Cotman and Zofija Mazej Kukovic. Mitteilung an die presse - 2876. tagung des rates beschäftigung, sozialpolitik, gesundheit und verbraucherschutz. June 2008. URL: http://www.consilium.europa.eu/ueDocs/cms_Data/docs/pressData/de/lsa/101752.pdf.

[Coo90] Gregory F. Cooper. The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence*, 42(2):393–405, 1990. ISSN: 0004-3702. DOI: 10.1016/0004-3702(90)90060-D.

[CP11] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, May 2011. ISSN: 1573-7683. DOI: 10.1007/s10851-010-0251-1.

[CS90] Susan Crawford and Loretta Stucki. Peer review and the changing research record. *Journal of the American Society for Information Science*, 41(3):223–228, 1990. DOI: 10.1002/(SICI)1097-4571(199004)41:3<223::AID-ASI14>3.0.CO;2-3.

[CVS+13] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, December 2013. ISSN: 1618-727X. DOI: 10.1007/s10278-013-9622-7.

[CZC+18] Hu Chen, Yi Zhang, Yunjin Chen, Junfeng Zhang, Weihua Zhang, Huaiqiang Sun, Yang Lv, Peixi Liao, Jiliu Zhou, and Ge Wang. Learn: learned experts' assessment-based reconstruction network for sparse-data ct. *IEEE Transactions on Medical Imaging*, 37(6):1333–1347, 2018. DOI: 10.1109/TMI.2018.2805692.

[DCM+14] D.R. Dance, S. Christofides, A.D.A. Maidment, I.D. McLean, and K.H. Ng. *Diagnostic Radiology Physics*. Non-serial Publications. INTERNATIONAL ATOMIC ENERGY AGENCY, Vienna, 2014. ISBN: 978-92-0-131010-1. URL: https://www.iaea.org/publications/8841/diagnostic-radiology-physics.

[Del34] Boris Nikolayevich Delaunay. Sur la sphère vide. *Bulletin of Academy of Sciences of the USSR. 7*, (6):793–800, 1934.

[DNS+08] Frank Dennerlein, Frédéric Noo, Harald Schöndube, Günter Lauritsch, and Joachim Hornegger. A factorization approach for cone-beam reconstruction on a circular short-scan. *IEEE transactions on medical imaging*, 27(7):887–896, 2008. ISSN: 1558-254X. DOI: 10.1109/TMI.2008.922705.

[DUS+17] Bram Duyx, Miriam J.E. Urlings, Gerard M.H. Swaen, Lex M. Bouter, and Maurice P. Zeegers. Scientific citations favor positive results: a systematic review and meta-analysis. *Journal of Clinical Epidemiology*, 88:92–101, 2017. ISSN: 0895-4356. DOI: 10.1016/j.jclinepi.2017.06.002.

[EF99] H Erdogan and J A Fessler. Ordered subsets algorithms for transmission tomography. *Physics in Medicine and Biology*, 44(11):2835–2851, 1999. DOI: 10.1088/0031-9155/44/11/311.

[Fal+19] William Falcon et al. Pytorch lightning. *GitHub*, 3, 2019. URL: https://github.com/PyTorchLightning/pytorch-lightning.

[FDK84] L. A. Feldkamp, L. C. Davis, and J. W. Kress. Practical cone-beam algorithm. *J. Opt. Soc. Am. A*, 1(6):612–619, June 1984. DOI: 10.1364/JOSAA.1.000612.

[FMB+06] Thomas G. Flohr, Cynthia H. McCollough, Herbert Bruder, Martin Petersilka, Klaus Gruber, Christoph Süβ, Michael Grasruck, Karl Stierstorfer, Bernhard Krauss, Rainer Raupach, Andrew N. Primak, Axel Küttner, Stefan Achenbach, Christoph Becker, Andreas Kopp, and Bernd M. Ohnesorge. First performance evaluation of a dual-source ct (dsct) system. *European Radiology*, 16(2):256–268, February 2006. ISSN: 1432-1084. DOI: 10.1007/s00330-005-2919-2.

[Fuk80] Kunihiko Fukushima. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, April 1980. ISSN: 1432-0770. DOI: 10.1007/BF00344251.

[GBH70] Richard Gordon, Robert Bender, and Gabor T. Herman. Algebraic reconstruction techniques (art) for three-dimensional electron microscopy and x-ray photography. *Journal of Theoretical Biology*, 29(3):471–481, 1970. ISSN: 0022-5193. DOI: 10.1016/0022-5193(70)90109-8. URL: http://www.sciencedirect.com/science/article/pii/0022519370901098.

[GBS+15] Olya Grove, Anders E. Berglund, Matthew B. Schabath, Hugo J.W.L. Aerts, Andre Dekker, Hua Wang, Emmanuel Rios Velazquez, Philippe Lambin, Yuhua Gu, Yoganand Balagurunathan, Edward Eikman, Robert A. Gatenby, Steven Eschrich, and Robert J. Gillies. Data from: quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma, 2015. DOI: 10.7937/K9/TCIA.2015.A6V7JIWX. URL: https://wiki.cancerimagingarchive.net/x/8IUiAQ.

[Gev06] Tal Geva. Magnetic resonance imaging: historical perspective. *Journal of Cardiovascular Magnetic Resonance*, 8(4):573–580, August 2006. DOI: 10.1080/10976640600755302.

[Gil72] Peter Gilbert. Iterative methods for the three-dimensional reconstruction of an object from projections. *Journal of Theoretical Biology*, 36(1):105–117, 1972. ISSN: 0022-5193. DOI: 10.1016/0022-5193(72)90180-4.

[GKR+21] Suhita Ghosh, Andreas Krug, Georg Rose, and Sebastian Stober. Perception-aware losses facilitate ct denoising and artifact removal. In *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*, pages 1–6, 2021. DOI: 10.1109/ICHMS53169.2021.9582444.

[Gre55] C.V. Gregg. 2483. the quotient of two quadratic functions. *The Mathematical Gazette*, 39(327):50–52, February 1955. DOI: 10.2307/3611091.

[HG08] Quan Huynh-Thu and Mohammad Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics Letters*, 44:800–801(1), 13, June 2008. ISSN: 0013-5194. DOI: 10.1049/el:20080522.

[HHL+16] Sung-Yuan Hu, Ming-Shun Hsieh, Meng-Yu Lin, Chiann-Yi Hsu, Tzu-Chieh Lin, Chorng-Kuang How, Chen-Yu Wang, Jeffrey Che-Hung Tsai, Yu-Hui Wu, and Yan-Zin Chang. Trends of ct utilisation in an emergency department in taiwan: a 5-year retrospective study. *BMJ Open*, 6(6), 2016. ISSN: 2044-6055. DOI: 10.1136/bmjopen-2015-010973. eprint: https://bmjopen.bmj.com/content/6/6/e010973.full.pdf. URL: https://bmjopen.bmj.com/content/6/6/e010973.

[HKL+18] Chang Min Hyun, Hwa Pyung Kim, Sung Min Lee, Sungchul Lee, and Jin Keun Seo. Deep learning for undersampled MRI reconstruction. *Physics in Medicine and Biology*, 63(13):135007, June 2018. DOI: 10.1088/1361-6560/aac71a.

[HL89] T. Hebert and R. Leahy. A generalized em algorithm for 3-d bayesian reconstruction from poisson data using gibbs priors. *IEEE Transactions on Medical Imaging*, 8(2):194–202, 1989. DOI: 10.1109/42.24868.

[HNL07] Telle Hailikari, Anne Nevgi, and Sari Lindblom-Ylänne. Exploring alternative ways of assessing prior knowledge, its components and their relation to student achievement: a mathematics based case study. *Studies in Educational Evaluation*, 33(3):320–337, 2007. ISSN: 0191-491X. DOI: 10.1016/j.stueduc.2007.07.007. URL: https://www.sciencedirect.com/science/article/pii/S0191491X07000351.

[Hou73] G. N. Hounsfield. Computerized transverse axial scanning (tomography): part 1. description of system. *The British Journal of Radiology*, 46(552):1016–1022, 1973. DOI: 10.1259/0007-1285-46-552-1016.

[HSW89] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. ISSN: 0893-6080. DOI: 10.1016/0893-6080(89)90020-8.

[HSY20] Yoseob Han, Leonard Sunwoo, and Jong Chul Ye. K-space deep learning for accelerated mri. *IEEE Transactions on Medical Imaging*, 39(2):377–386, 2020. DOI: 10.1109/TMI.2019.2927101.

[HXL+11] Yining Hu, Lizhe Xie, Limin Luo, Jean Claude Nunes, and Christine Toumoulin. L0 constrained sparse reconstruction for multi-slice helical CT reconstruction. *Physics in Medicine and Biology*, 56(4):1173–1189, February 2011. DOI: 10.1088/0031-9155/56/4/018.

[JFS+13] Saleem Jahangeer, Patrick Forde, Declan Soden, and John Hinchion. Review of current thermal ablation treatment for lung cancer and the potential of electrochemotherapy as a means for treatment of lung tumours. *Cancer Treatment Reviews*, 39(8):862–871, 2013. ISSN: 0305-7372. DOI: 10.1016/j.ctrv.2013.03.007. URL: https://www.sciencedirect.com/science/article/pii/S0305737213000704.

[JMF+17] Kyong Hwan Jin, Michael T. McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017. DOI: 10.1109/TIP.2017.2713099.

[Joh38] Fritz John. The ultrahyperbolic differential equation with four independent variables. *Duke Mathematical Journal*, 4(2):300–322, 1938. DOI: 10.1215/S0012-7094-38-00423-5.

[KCW+18] Shaan Khurshid, Seung Hoan Choi, Lu-Chen Weng, Elizabeth Y. Wang, Ludovic Trinquart, Emelia J. Benjamin, Patrick T. Ellinor, and Steven A. Lubitz. Frequency of cardiac rhythm abnormalities in a half million adults. *Circulation: Arrhythmia and Electrophysiology*, 11(7):e006273, 2018. DOI: 10.1161/CIRCEP.118.006273.

[KD88] David Klahr and Kevin Dunbar. Dual space search during scientific reasoning. *Cognitive Science*, 12(1):1–48, 1988. ISSN: 0364-0213. DOI: https://doi.org/10.1016/0364-0213(88)90007-9.

[KM75] R.L. Kashyap and M.C. Mittal. Picture reconstruction from projections. *IEEE Transactions on Computers*, C-24(9):915–923, 1975. DOI: 10.1109/T-C.1975.224337.

[KS88] A. C. Kak and Malcolm Slaney. *Principles of Computerized Tomographic Imaging*. IEEE Press, 1988.

[KSK+90] W A Kalender, W Seissler, E Klotz, and P Vock. Spiral volumetric ct with single-breath-hold technique, continuous transport, and continuous scanner rotation. *Radiology*, 176(1):181–183, 1990. DOI: 10.1148/radiology.176.1.2353088.

[Lan86] Serge Lang. *Determinants*. In *Introduction to Linear Algebra*. Springer New York, New York, NY, 1986, pages 195–232. ISBN: 978-1-4612-1070-2. DOI: 10.1007/978-1-4612-1070-2_7.

[LAU73] P. C. LAUTERBUR. Image formation by induced local interactions: examples employing nuclear magnetic resonance. *Nature*, 242(5394):190–191, March 1973. DOI: 10.1038/242190a0.

[LB09] Chaofeng Li and Alan C. Bovik. Three-component weighted structural similarity index. In Susan P. Farnand and Frans Gaykema, editors, *Image Quality and System Performance VI*, volume 7242, pages 252–260. International Society for Optics and Photonics, SPIE, 2009. DOI: 10.1117/12.811821.

[LBD+89] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. DOI: 10.1162/neco.1989.1.4.541.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. ISSN: 1476-4687. DOI: 10.1038/nature14539.

[LBV06] P.J. La Riviere, Junguo Bian, and P.A. Vargas. Penalized-likelihood sinogram restoration for computed tomography. *IEEE Transactions on Medical Imaging*, 25(8):1022–1036, 2006. DOI: 10.1109/TMI.2006.875429.

[LFB10] Yong Long, Jeffrey A. Fessler, and James M. Balter. 3d forward and back-projection for x-ray ct using separable footprints. *IEEE Transactions on Medical Imaging*, 29(11):1839–1850, 2010. DOI: 10.1109/TMI.2010.2050898.

[LFM+15] Carlo Liguori, Giulia Frauenfelder, Carlo Massaroni, Paola Saccomandi, Francesco Giurazza, Francesca Pitocco, Riccardo Marano, and Emiliano Schena. Emerging clinical applications of computed tomography. eng. *Medical devices (Auckland, N.Z.)*, 8:265–278, June 2015. ISSN: 1179-1470. DOI: 10.2147/MDER.S70630.

[Lin18] Jyh-Miin Lin. Python non-uniform fast fourier transform (pynufft): an accelerated non-cartesian mri package on a heterogeneous platform (cpu/gpu). *Journal of Imaging*, 4(3):51, 2018.

[Lin70] Seppo Linnainmaa. *Algoritmin Kumulatiivinen Pyöristysvirhe Yksittäisten Pyöristysvirheiden Taylor-Kehitelmänä*. In Finnish. University of Helsinki, 1970. DOI: 10.1007/BF01931367. English version published in 1976.

[LLK+19] Hoyeon Lee, Jongha Lee, Hyeongseok Kim, Byungchul Cho, and Seungryong Cho. Deep-neural-network-based sinogram synthesis for sparse-view ct image reconstruction. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 3(2):109–119, 2019. DOI: 10.1109/TRPMS.2018.2867611.

[LSD15] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, Los Alamitos, CA, USA. IEEE Computer Society, June 2015. DOI: 10.1109/CVPR.2015.7298965.

[LSG+21] Johannes Leuschner, Maximilian Schmidt, Poulami Somanya Ganguly, Vladyslav Andriiashen, Sophia Bethany Coban, Alexander Denker, Dominik Bauer, Amir Hadjifaradji, Kees Joost Batenburg, Peter Maass, and Maureen van Eijnatten. Quantitative comparison of deep learning-based image reconstruction methods for low-dose and sparse-angle ct applications. *Journal of Imaging*, 7(3), 2021. ISSN: 2313-433X. DOI: 10.3390/jimaging7030044. URL: https://www.mdpi.com/2313-433X/7/3/44.

[LTH+17] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, 2017. DOI: 10.1109/CVPR.2017.19.

[LZL+17] Kisuk Lee, Jonathan Zung, Peter Li, Viren Jain, and H. Sebastian Seung. Superhuman accuracy on the SNEMI3D connectomics challenge. *arXiv preprint arXiv:1706.00120*, 2017.

[Mar06] James E. Martin. *Physics for Radiation Protection: A Handbook*. Wiley, 2006. ISBN: 9783527406111.

[MAV17] Siddharth Mahendran, Haider Ali, and René Vidal. 3d pose regression using convolutional neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 494–495, 2017. DOI: 10.1109/CVPRW.2017.73.

[May22] Mayo Clinic Staff. Cardiac ablation. 2022. URL: https://www.mayoclinic.org/tests-procedures/cardiac-ablation/about/pac-20384993 (visited on 02/23/2022).

[MCH+20] CH McCollough, B Chen, DI Holmes, X Duan, Z Yu, L Xu, S Leng, and J Fletcher. Low dose CT image and projection data [data set]. *The Cancer Imaging Archive*, 2020. DOI: 10.7937/9NPB-2637.

[McK98] M H McKetty. The aapm/rsna physics tutorial for residents. x-ray attenuation. *RadioGraphics*, 18(1):151–163, 1998. DOI: 10.1148/radiographics.18.1.9460114. PMID: 9460114.

[MD07] Dale Miles and Robert Danforth. A clinician's guide to understanding cone beam volumetric imaging (cbvi). *Acad Dent Ther Stomatol*:1–13, January 2007.

[MHN+13] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Citeseer, 2013.

[MNA+17] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv:1710.03740*, 2017.

[MNA16] F. Milletari, N. Navab, and S. Ahmadi. V-net: fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571, Los Alamitos, CA, USA. IEEE Computer Society, October 2016. DOI: 10.1109/3DV.2016.79.

[MP43] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, December 1943. ISSN: 1522-9602. DOI: 10.1007/BF02478259.

[MP69] M. Minsky and S. Papert. *Perceptrons; an Introduction to Computational Geometry*. MIT Press, 1969. ISBN: 9780262630221.

[MPT+98] P. Mozzo, C. Procacci, A. Tacconi, P. Tinazzi Martini, and I. A. Bergamo Andreis. A new volumetric ct machine for dental imaging based on the cone-beam technique: preliminary results. *European Radiology*, 8(9):1558–1564, November 1998. DOI: 10.1007/s003300050586.

[MR06] Tarun K. Mittal and Michael B. Rubens. *Computed tomography techniques and principles. part a. electron beam computed tomography*. In *Noninvasive Imaging of Myocardial Ischemia*. Constantinos D. Anagnostopoulos, Petros Nihoyannopoulos, Jeroen J. Bax, and Ernst van der Wall, editors. Springer London, London, 2006, pages 93–98. ISBN: 978-1-84628-156-3. DOI: 10.1007/1-84628-156-3_6.

[MW47] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*:50–60, 1947.

[NHD+07] Frédéric Noo, Stefan Hoppe, Frank Dennerlein, Günter Lauritsch, and Joachim Hornegger. A new scheme for view-dependent data differentiation in fan-beam and cone-beam computed tomography. *Physics in Medicine and Biology*, 52(17):5393–5414, August 2007. DOI: 10.1088/0031-9155/52/17/020.

[NWV+15] Masih Nilchian, John Paul Ward, Cédric Vonesch, and Michael Unser. Optimized kaiser–bessel window functions for computed tomography. *IEEE Transactions on Image Processing*, 24(11):3826–3833, 2015. DOI: 10.1109/TIP.2015.2451955.

[OWK09] Robert C. Orth, Michael J. Wallace, and Michael D. Kuo. C-arm cone-beam ct: general principles and technical considerations for use in interventional radiology. *Journal of Vascular and Interventional Radiology*, 20(7, Supplement):S538–S544, 2009. ISSN: 1051-0443. DOI: https://doi.org/10.1016/j.jvir.2009.04.026.

[PFB+19] Tim Pfeiffer, Robert Frysch, Richard N. K. Bismark, and Georg Rose. CTL: modular open-source C++-library for CT-simulations. In Samuel Matej and Scott D. Metzler, editors, *15th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine*, volume 11072, pages 269–273. International Society for Optics and Photonics, SPIE, 2019. DOI: 10.1117/12.2534517.

[PGM+19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: an imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[PRC+22] Kumari Pooja, Zaccharie Ramzi, G.R. Chaithya, and Philippe Ciuciu. Mc-pdnet: deep unrolled neural network for multi-contrast mr image reconstruction from undersampled k-space data. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2022. DOI: 10.1109/ISBI52829.2022.9761583.

[Rad17] J. Radon. Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. *Berichte über die Verhandlungen der Sächsische Akademie der Wissenschaften*, 69:262–277, 1917.

[RB08] Justus Randolph and Roman Bednarik. Publication bias in the computer science education research literature. *J. UCS*, 14:575–589, January 2008.

[RFB15a] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: convolutional networks for biomedical image segmentation. In *Med Imag Comput Comput Assis Interv*, 2015.

[RFB15b] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing, 2015.

[rhtt] rahnema1 (https://stackoverflow.com/users/6579744/rahnema1). How to interpolate using in polar coordinate. Stack Overflow. eprint: `https://stackoverflow.com/questions/40858534`. URL: `https://stackoverflow.com/questions/40858534`. version: 2016-11-29.

[RLS+15] Holger R Roth, Le Lu, Ari Seff, Kevin M Cherry, Joanne Hoffman, Shijun Wang, Jiamin Liu, Evrim Turkbey, and Ronald M Summers. A new 2.5 d representation for lymph node detection in ct. the cancer imaging archive, 2015. DOI: `10.7937/K9/TCIA.2015.AQIIDCNM`.

[Ron20] Matteo Ronchetti. Torchradon: fast differentiable routines for computed tomography. *arXiv preprint arXiv:2009.14788*, 2020. eprint: `arXiv:2009.14788`.

[Ros57] F. Rosenblatt. The perceptron - A perceiving and recognizing automaton. Technical report 85-460-1, Cornell Aeronautical Laboratory, Ithaca, New York, January 1957.

[RPI77] RA Rutherford, BR Pullan, and I Isherwood. The physical performance of a prototype ct5000 emi body scanner. In *The First European Seminar on Computerised Axial Tomography in Clinical Practice*, pages 301–311. Springer, 1977.

[SFS+21] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021. DOI: `https://doi.org/10.3322/caac.21660`.

[Sid85] Robert L. Siddon. Fast calculation of the exact radiological path for a three-dimensional ct array. *Medical Physics*, 12(2):252–255, 1985. DOI: `10.1118/1.595715`. URL: `https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.595715`.

[SMB10] Dominik Scherer, Andreas Müller, and Sven Behnke. Evaluation of pooling operations in convolutional architectures for object recognition. In Konstantinos Diamantaras, Wlodek Duch, and Lazaros S. Iliadis, editors, *Artificial Neural Networks – ICANN 2010*, pages 92–101, Berlin, Heidelberg. Springer Berlin Heidelberg, 2010. ISBN: 978-3-642-15825-4.

[SP08] Emil Y Sidky and Xiaochuan Pan. Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. *Physics in Medicine and Biology*, 53(17):4777–4807, August 2008. DOI: 10.1088/0031-9155/53/17/021.

[SZ15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL: http://arxiv.org/abs/1409.1556.

[TBS+06] Jean-Baptiste Thibault, Charles A. Bouman, Ken D. Sauer, and Jiang Hsieh. A recursive filter for noise reduction in statistical iterative tomographic imaging. In Charles A. Bouman, Eric L. Miller, and Ilya Pollak, editors, *Computational Imaging IV*, volume 6065, pages 264–273. International Society for Optics and Photonics, SPIE, 2006. DOI: 10.1117/12.660281.

[TBZ+18] Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, Joao Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J Pal. Deep complex networks. In *International Conference on Learning Representations*, 2018. URL: https://openreview.net/forum?id=H1T2hmZAb.

[Tuy83] Heang K. Tuy. An inversion formula for cone-beam reconstruction. *SIAM Journal on Applied Mathematics*, 43(3):546–552, 1983. DOI: 10.1137/0143035.

[vAPC+16] Wim van Aarle, Willem Jan Palenstijn, Jeroen Cant, Eline Janssens, Folkert Bleichrodt, Andrei Dabravolski, Jan De Beenhouwer, K. Joost Batenburg, and Jan Sijbers. Fast and flexible x-ray tomography using the astra toolbox. *Opt. Express*, 24(22):25129–25147, October 2016. DOI: 10.1364/OE.24.025129. URL: http://opg.optica.org/oe/abstract.cfm?URI=oe-24-22-25129.

[WB09] Zhou Wang and Alan C Bovik. Mean squared error: love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1):98–117, 2009.

[WBS+04] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. DOI: 10.1109/TIP.2003.819861.

[Wer82] Paul J. Werbos. Applications of advances in nonlinear sensitivity analysis. In R. F. Drenick and F. Kozin, editors, *System Modeling and Optimization*, pages 762–770, Berlin, Heidelberg. Springer Berlin Heidelberg, 1982.

[WGC+16] Tobias Würfl, Florin C. Ghesu, Vincent Christlein, and Andreas Maier. Deep learning computed tomography. In Sebastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gozde Unal, and William Wells, editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016*, pages 432–440, Cham. Springer International Publishing, 2016.

[WHC+18] Tobias Würfl, Mathis Hoffmann, Vincent Christlein, Katharina Breininger, Yixin Huang, Mathias Unberath, and Andreas K. Maier. Deep learning computed tomography: learning projection-domain weights from image domain in limited angle problems. *IEEE Transactions on Medical Imaging*, 37(6):1454–1463, 2018. DOI: 10.1109/TMI.2018.2833499.

[WMM+20] Ralph C. Wang, Diana L. Miglioretti, Emily C. Marlow, Marilyn L. Kwan, May K. Theis, Erin J. A. Bowles, Robert T. Greenlee, Alanna K. Rahm, Natasha K. Stout, Sheila Weinmann, and Rebecca Smith-Bindman. Trends in imaging for suspected pulmonary embolism across us health care systems, 2004 to 2016. *JAMA Network Open*, 3(11):e2026930–e2026930, November 2020. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2020.26930.

[WSB03] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, 1398–1402 Vol.2, 2003. DOI: 10.1109/ACSSC.2003.1292216.

[WT97] Brian Wandell and Stephen Thomas. Foundations of vision. *Psyccritiques*, 42(7), 1997.

[XYM+12] Qiong Xu, Hengyong Yu, Xuanqin Mou, Lei Zhang, Jiang Hsieh, and Ge Wang. Low-dose x-ray ct reconstruction via dictionary learning. *IEEE Transactions on Medical Imaging*, 31(9):1682–1697, 2012. DOI: 10.1109/TMI.2012.2195669.

[Zam89] L. Zambresky. A verification study of the global wam model december 1987 - november 1988. (63):86, May 1989. URL: https://www.ecmwf.int/node/13201.

[ZD20] Hai-Miao Zhang and Bin Dong. A review on deep learning in medical image reconstruction. *Journal of the Operations Research Society of China*:1–30, 2020.

[ZGF+17] H. Zhao, O. Gallo, I. Frosio, and J. Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2017. DOI: 10.1109/TCI.2016.2644865.

[ZH95a] S.-R. Zhao and H. Halling. A new fourier method for fan beam reconstruction. In *1995 IEEE Nuclear Science Symposium and Medical Imaging Conference Record*, volume 2, 1287–1291 vol.2, 1995. DOI: 10.1109/NSSMIC.1995.510494.

[ZH95b] Shuang-Ren Zhao and Horst Halling. Reconstruction of cone beam projections with free source path by a generalized fourier method. In *Proceedings of the 1995 International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine*, pages 323–327, 1995.

[Zha04] Harry Zhang. The optimality of naive bayes. In Valerie Barr and Zdravko Markov, editors, *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)* (Miami Beach, Florida, USA). AAAI Press, 2004.

[ZHL+16] Hao Zhang, Hao Han, Zhengrong Liang, Yifan Hu, Yan Liu, William Moore, Jianhua Ma, and Hongbing Lu. Extracting information from previous full-dose ct scan for knowledge-based bayesian reconstruction of current low-dose ct images. *IEEE Transactions on Medical Imaging*, 35(3):860–870, 2016. DOI: 10.1109/TMI.2015.2498148.

[ZHM+14] Hua Zhang, Jing Huang, Jianhua Ma, Zhaoying Bian, Qianjin Feng, Hongbing Lu, Zhengrong Liang, and Wufan Chen. Iterative reconstruction for x-ray computed tomography using prior-image induced nonlocal regularization. *IEEE Transactions on Biomedical Engineering*, 61(9):2367–2378, 2014. DOI: 10.1109/TBME.2013.2287244.

[Zim00] Corinne Zimmerman. The development of scientific reasoning skills. *Developmental Review*, 20(1):99–149, 2000. ISSN: 0273-2297. DOI: https://doi.org/10.1006/drev.1999.0497.

[ZLC+18] Bo Zhu, Jeremiah Z. Liu, Stephen F. Cauley, Bruce R. Rosen, and Matthew S. Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487–492, March 2018. ISSN: 1476-4687. DOI: 10.1038/nature25988.

[ZNG08] Andy Ziegler, Tim Nielsen, and Michael Grass. Iterative reconstruction of a region of interest for transmission tomography. *Medical Physics*, 35(4):1317–1327, 2008. DOI: 10.1118/1.2870219. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.2870219.

[ZWZ+18] Hao Zhang, Jing Wang, Dong Zeng, Xi Tao, and Jianhua Ma. Regularization strategies in statistical image reconstruction of low-dose x-ray ct: a review. *Medical Physics*, 45(10):e886–e907, 2018. DOI: 10.1002/mp.13123. eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.13123. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.13123.

[ZZH+13] Hua Zhang, Shanli Zhang, Debin Hu, Dong Zeng, Zhaoying Bian, Lijun Lu, Jianhua Ma, and Jing Huang. Threshold choices of huber regularization using global- and local-edge-detecting operators for x-ray computed tomographic reconstruction. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2352–2355, 2013. DOI: 10.1109/EMBC.2013.6610010.

Author's Contributions

[DED+21] Max Dünnwald, Philipp Ernst, Emrah Düzel, Klaus Tönnies, Matthew J. Betts, and Steffen Oeltze-Jafra. Fully automated deep learning-based localization and segmentation of the locus coeruleus in aging and parkinson's disease using neuromelanin-sensitive mri. *International Journal of Computer Assisted Radiology and Surgery*, 16(12):2129–2135, December 2021. ISSN: 1861-6429. DOI: 10.1007/s11548-021-02528-5.

[ECR+21] Philipp Ernst, Soumick Chatterjee, Georg Rose, Oliver Speck, and Andreas Nürnberger. Sinogram upsampling using Primal-Dual UNet for undersampled CT and radial MRI reconstruction. *arXiv e-prints*:arXiv:2112.13443, arXiv:2112.13443, December 2021. arXiv: 2112.13443 [eess.IV].

[ECR+22a] Philipp Ernst, Soumick Chatterjee, Georg Rose, and Andreas Nürnberger. Primal-dual UNet for sparse view cone beam computed tomography volume reconstruction. In *Medical Imaging with Deep Learning*, 2022. URL: https://openreview.net/forum?id=RVKcDeJ2fCi.

[ECR+22b] Philipp Ernst, Soumick Chatterjee, Georg Rose, Oliver Speck, and Andreas Nürnberger. Sinogram upsampling using primal-dual unet for undersampled ct and radial mri reconstruction. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 2022.

[EGR+22] Philipp Ernst, Suhita Ghosh, Georg Rose, and Andreas Nürnberger. Dual branch prior-segnet: CNN for interventional CBCT using planning scan and auxiliary segmentation loss. In *Medical Imaging with Deep Learning*, 2022. URL: https://openreview.net/forum?id=uhv14tCLmoB.

[EHH+19] Philipp Ernst, Georg Hille, Christian Hansen, Klaus Tönnies, and Marko Rak. A cnn-based framework for statistical assessment of spinal shape and curvature in whole-body mri images of large populations. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 3–11, Cham. Springer International Publishing, 2019.

[ERH+21] Philipp Ernst, Marko Rak, Christian Hansen, Georg Rose, and Andreas Nürnberger. Trajectory upsampling for sparse conebeam projections using convolutional neural networks. In Georg Schramm, Ahmadreza Rezaei, Kris Thielemans, and Johan Nuyts, editors, *Proceedings of the 16th Virtual International Meeting on Fully 3D Image Reconstruction in Radiology and Nuclear Medicine*, pages 285–288, 2021.

[ERN19] Philipp Ernst, Georg Rose, and Andreas Nürnberger. Comparison of optimization methods for few view ct using deep learning. Poster presented at: 4th Image-Guided Interventions Conference: Digitalization in Medicine, Mannheim, Germany, November 2019.

[ERN21] Philipp Ernst, Georg Rose, and Andreas Nürnberger. Sparse view deep differentiated backprojection for circular trajectories in cbct. In Georg Schramm, Ahmadreza Rezaei, Kris Thielemans, and Johan Nuyts, editors, *Proceedings of the 16th Virtual International Meeting on Fully 3D Image Reconstruction in Radiology and Nuclear Medicine*, pages 463–466, 2021.

[GER+22] Suhita Ghosh, Philipp Ernst, Georg Rose, Andreas Nürnberger, and Sebastian Stober. Towards patient specific reconstruction using perception-aware cnn and planning ct as prior. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2022. DOI: 10.1109/ISBI52829.2022.9761462.

[KGB+21] A. Emre Kavur, N. Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, Bora Baydar, Dmitry Lachinov, Shuo Han, Josef Pauli, Fabian Isensee, Matthias Perkonigg, Rachana Sathish, Ronnie Rajan, Debdoot Sheet, Gurbandurdy Dovletov, Oliver Speck, Andreas Nürnberger, Klaus H. Maier-Hein, Gözde Bozdağı Akar, Gözde Ünal, Oğuz Dicle, and M. Alper Selver. CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, April 2021. ISSN: 1361-8415. DOI: 10.1016/j.media.2020.101950.

[XEL+21] Jiahua Xu, Philipp Ernst, Tung Lung Liu, and Andreas Nürnberger. Dual skip connections minimize the false positive rate of lung nodule detection in ct images. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 3217–3220, 2021. DOI: 10.1109/EMBC46164.2021.9630096.

Appendix A

Primal-Dual UNet for Undersampled Radial MRI

As described in Sec. 5.2, the results that are shown in this chapter only focus on MRI data. They are included here to show that the methods that were developed in this thesis have not only been designed for a very specific use case, but are even applicable for rather different imaging modalities when converted to data that is mathematically similar to CT projections.

A.1 Dataset

Two different publicly available benchmark data sets of two different organs were used in this research: IXI data set[1] for brain and CHAOS challenge data set [KGB+21] for abdomen. The IXI data set contains nearly 600 brain MRIs of normal healthy subjects, acquired using different MRI protocols (T1w, T2w, PDw, MRA, and DWI), collected from three different hospitals at two different field strengths (1.5T and 3T). 30 central slices from 100, 35 and 50 T1w volumes acquired at 3T were used in this study as the training, validation and test set, respectively. The CHAOS challenge data set contains abdominal MRIs of 40 healthy subjects, acquired using two different MR sequences: T1-Dual (In-phase and Opposed-phase) and T2-SPIR. All the slices from 24, 6 and 10 subjects (three volumes each: T1-in, T1-opposed, and T2) were used in the training, validation, and test set, respectively. All the images were interpolated with sinc interpolation to have an in-plane matrix size of 256x256.

The data sets do not contain any raw MRI data, only the magnitude images, which were treated as the fully-sampled groundtruth. The corresponding single-coil radial k-spaces of those magnitude images were generated using NUFFT (implemented in PyNUFFT [Lin18]). The fully-sampled raw data was considered to have the number of spokes (radial acquisitions) as twice the base resolution, which was 512 for this data set. The sampling was performed following the equidistant radial sampling scheme,

[1] IXI Dataset: `https://brain-development.org/ixi-dataset/`.

where the angle between the spokes, calculated as $\Delta\phi = \pi \div n_{Sp}$ for $n_{Sp} \in \mathbf{N}$ spokes, was $\Delta\phi = 0.351\,562\,5°$. A NUFFT was performed on each slice of the MRI volumes, following $\Delta\phi$, to obtain the equivalent fully-sampled radial k-space.

Pre-Processing The fully-sampled sinograms of the respective radial k-spaces were obtained by applying the one-dimensional inverse Fourier transform on each spoke of the k-space. To keep the setup similar to the one described in Sec. 5.2.5, the spokes were shifted by half a detector pixel using sinc interpolation and were cropped to the central 363 detector pixels. This step was necessary because each spoke extracted by NUFFT contained 512 frequency components, which corresponds to a detector pixel number of 512, as well. In contrast to the simulated CT projections in Sec. 5.2.5, the MR sinograms correspond to parallel-beam projections according to the Fourier slice theorem.

Undersampling To simulate the undersampled data sets, the sinograms were made sparse by retaining only every n^{th} projection (spoke), where n denotes the level of sparsity. Two levels were used: Sparse 8 and Sparse 16.

A.2 Results

The proposed Primal-Dual UNet was compared (see Sec. 5.2.2) against the undersampled radial k-space reconstruction using PyNUFFT [Lin18] (referred to as *Undersampled (NUFFT)*) by applying FBP on the corresponding sinograms (obtained by applying 1D-iFFT on each spoke, explained in Sec. A.1), referred to here as *Sinogram Bilinear*, and finally, against three deep learning baseline models: Reconstruction UNet [HKL$^+$18], Sinogram UNet [LLK$^+$19], and Learned Primal-Dual Network [AÖ18] - for two different publicly available benchmark data sets for two different organs: IXI for T1w brain MRIs and CHAOS for T1-Dual and T2w abdominal MRIs. Experiments were performed for two different levels of undersamplings: with an acceleration factor of 8 and 16 - which, in terms of the sparsity of the corresponding sinograms for equidistant radial samplings (see Sec. A.1 and A.1), are referred to here as Sparse 8 and 16 - to have the same terminology for both CT and MRI.

A.2.1 IXI Dataset

Qualitative results of SSIM and RMSE are shown in Tab. A.1, whereas the range of the resultant SSIM values with the help of box plots is shown in Fig. A.1 for Sparse 8 and 16, respectively. It can be observed that the proposed model outperformed all the baseline methods in terms of both SSIM and RMSE, and the statistical tests revealed that these improvements were significant. In terms of average SSIM values, the Primal-Dual UNet achieved improvements of 1.9% and 4.15% over the main baseline Primal-Dual Network for Sparse 8 and 16, respectively. Qualitative comparisons of the results using difference images and SSIM maps are shown in Fig. A.2 for Sparse 8 for Sparse 16.

Table A.1: Resultant metrics for MRI for the IXI data set (mean±std)

Method	SSIM	
	Sparse 8	Sparse 16
Undersampled (NUFFT)	0.595±0.027	0.410±0.021
Bilinear Sinogram	0.819±0.026	0.682±0.035
Sinogram UNet	0.860±0.044	0.782±0.032
Reconstruction UNet	0.948±0.011	0.877±0.025
Primal-Dual Network	0.947±0.012	0.867±0.025
Primal-Dual UNet	**0.965±0.008**	**0.903±0.019**

Method	RMSE	
	Sparse 8	Sparse 16
Undersampled (NUFFT)	0.046±0.009	0.085±0.016
Bilinear Sinogram	0.058±0.017	0.067±0.016
Sinogram UNet	0.058±0.021	0.073±0.023
Reconstruction UNet	0.021±0.006	0.037±0.012
Primal-Dual Network	0.021±0.006	0.041±0.015
Primal-Dual UNet	**0.017±0.005**	**0.034±0.011**

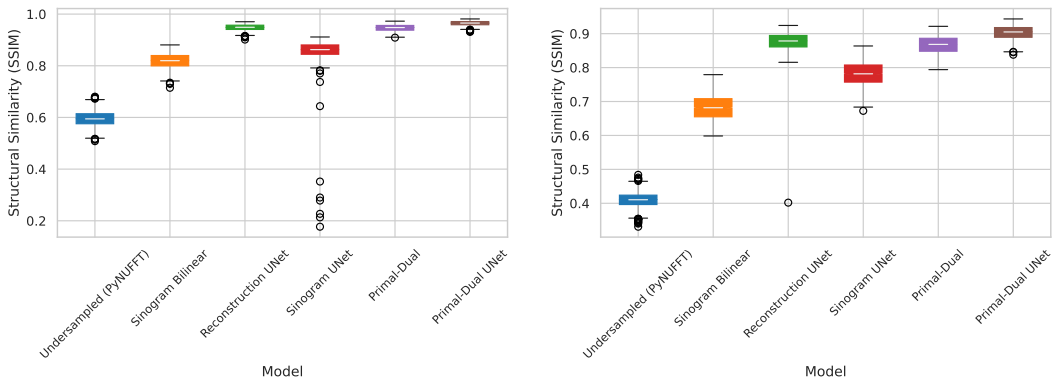


Figure A.1: Box-plots of the resultant SSIM values for MRI Sparse 8 (left) and Sparse 16 (right) for the IXI data set

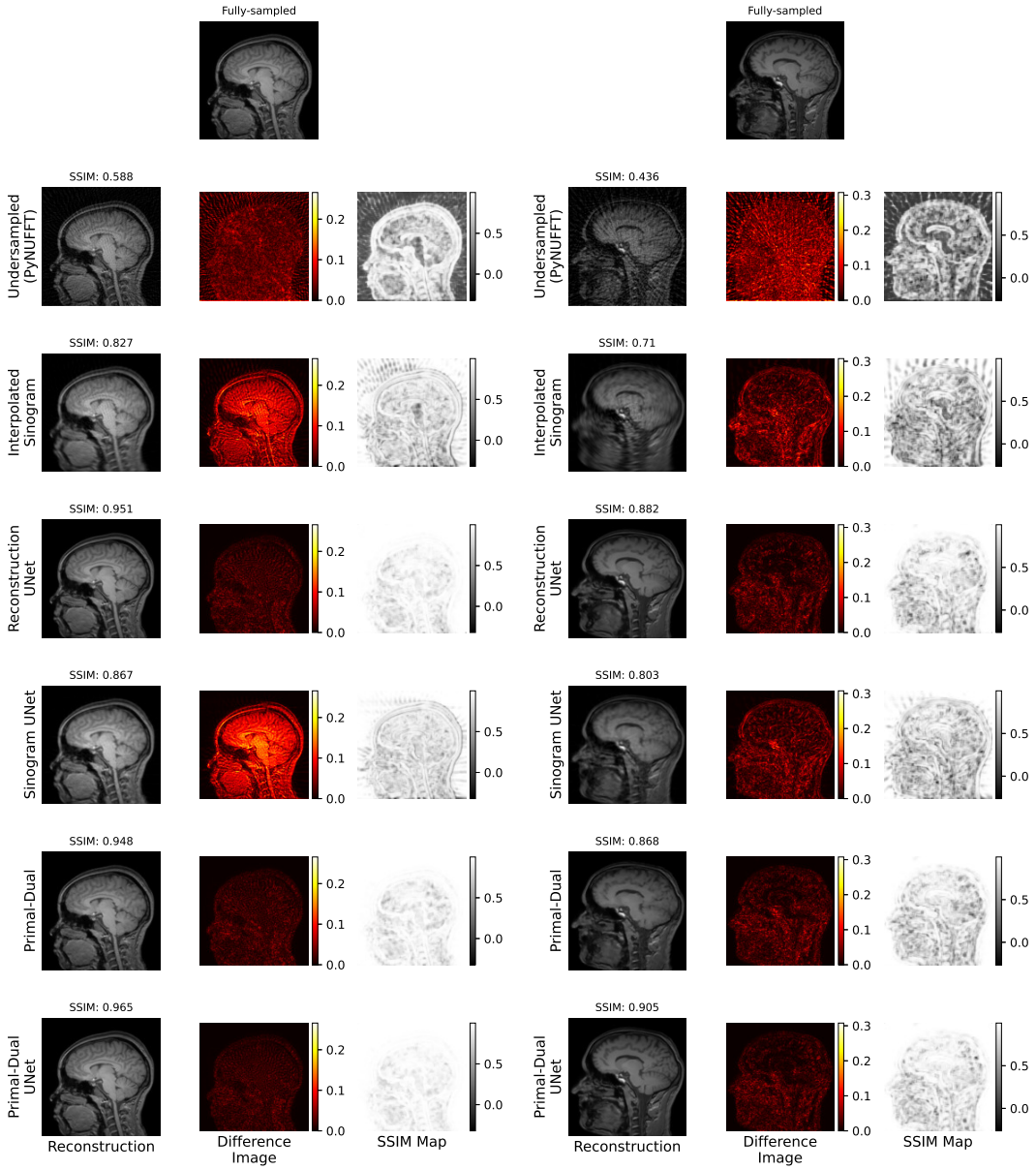


Figure A.2: Qualitative comparisons of the reconstructions of MRI Sparse 8 (left) and Sparse 16 (right) for the IXI data set

Table A.2: Resultant metrics for MRI for the CHAOS data set (mean±std)

Method	SSIM		RMSE	
	Sparse 8	Sparse 16	Sparse 8	Sparse 16
Undersampled (NUFFT)	0.528±0.092	0.359±0.060	0.034±0.015	0.068±0.027
Bilinear Sinogram	0.915±0.033	0.843±0.060	0.035±0.013	0.044±0.016
Sinogram UNet	0.926±0.026	0.897±0.038	0.032±0.011	0.036±0.014
Reconstruction UNet	0.981±0.013	0.943±0.029	0.011±0.007	0.022±0.014
Primal-Dual Network	0.982±0.012	0.949±0.025	0.010±0.005	0.021±0.010
Primal-Dual UNet	**0.986±0.011**	**0.957±0.023**	**0.009±0.006**	**0.018±0.011**

Table A.3: Resultant SSIM metrics for different types of MR acquisitions from the CHAOS data set (mean±std)

Method	Sparse 8			
	T2-SPIR	T1-Dual	T1 In-phase	T1 Opposed-phase
Undersampled (NUFFT)	0.599±0.060	0.493±0.084	0.430±0.058	0.557±0.053
Bilinear Sinogram	0.944±0.017	0.901±0.029	0.887±0.029	0.916±0.021
Sinogram UNet	0.950±0.014	0.915±0.023	0.904±0.023	0.925±0.017
Reconstruction UNet	0.983±0.009	0.980±0.015	0.975±0.014	0.985±0.013
Primal-Dual Network	0.983±0.008	0.981±0.013	0.976±0.013	0.986±0.011
Primal-Dual UNet	**0.987±0.007**	**0.985±0.012**	**0.981±0.011**	**0.989±0.011**
Method	Sparse 16			
	T2-SPIR	T1-Dual	T1 In-phase	T1 Opposed-phase
Undersampled (NUFFT)	0.394±0.051	0.342±0.057	0.306±0.043	0.378±0.045
Bilinear Sinogram	0.896±0.035	0.817±0.052	0.796±0.051	0.838±0.044
Sinogram UNet	0.927±0.021	0.882±0.036	0.867±0.036	0.898±0.027
Reconstruction UNet	0.950±0.019	0.939±0.033	0.926±0.031	0.952±0.029
Primal-Dual Network	0.955±0.017	0.945±0.028	0.934±0.028	0.957±0.022
Primal-Dual UNet	**0.961±0.015**	**0.955±0.026**	**0.945±0.026**	**0.964±0.023**

A.2.2 CHAOS Dataset

The final set of experiments was performed on the CHAOS MRI data set. Quantitatively, the proposed method outperformed all the baselines for both acceleration factors with statistical significance when being compared using SSIM and RMSE, reported in Tab. A.2. Fig. A.3 portrays the range of resultant SSIM values for Sparse 8 and 16, respectively. The proposed method improved the average SSIM values by 0.41% and 0.84% over the baseline Primal-Dual Network, respectively, for Sparse 8 and 16.

The CHAOS data set includes three different types of MRIs: T1 in-phase, T1 opposed-phase, and T2, acquired using two different sequences for two different contrasts: T1-Dual and T2-SPIR (explained in A.1) and all three were combined during training. The metrics that have been shown so far also included all these three types of MRIs together. However, in addition, they were also evaluated separately. Tab. A.3

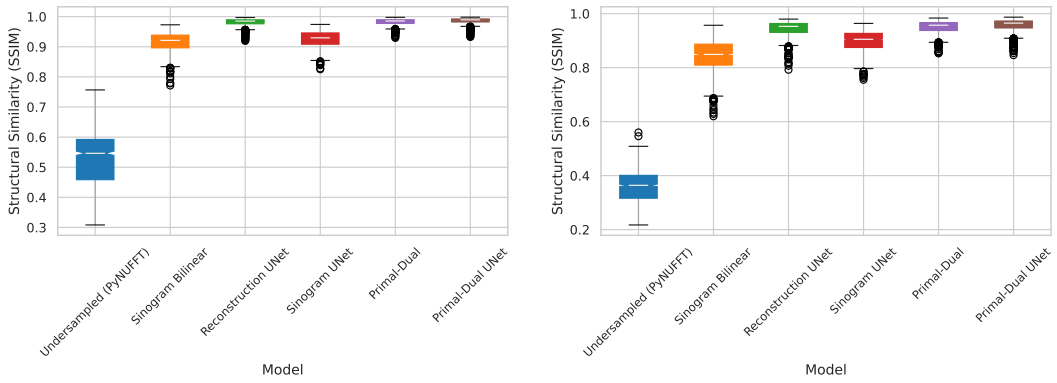


Figure A.3: Box-plots of the resultant SSIM values for MRI Sparse 8 (left) and Sparse 16 (right) for the CHAOS data set

shows the resultant SSIM values for each of the three types of MRIs separately, as well as for both types of acquisition sequences T1-Dual (all the results of T1 in-phase and opposed-phase) and T2-SPIR. The proposed method achieved statistically significant improvements over all the baselines in every scenario. Among the different types of MRIs, T1 opposed-phase achieved the best score, while T1 in-phase achieved the worst. When only the contrasts/sequences are taken into account, then T2-SPIR performed better than T1-Dual. These observations hold true for both levels of sparsity. However, in terms of the percentage of improvement of average SSIM values achieved by the proposed method against the primary baseline model (Primal-Dual Network), the amounts are different for the different levels of sparsity. Both T2-SPIR and T1-Dual improved 0.41% for Sparse 8, and for Sparse 16 obtained improvements of 0.63% and 1.06%, respectively. Considering the three types of MRIs separately, it is noteworthy that even though the T1 in-phase was the worst-performing, it managed to get the most amount of improvement 0.51% and 1.18% for Sparse 8 and 16. In contrast, the best-performing type of MRI secured the least amount of improvements: 0.30% and 0.73% for Sparse 8 and 16.

Finally, the reconstructions were compared qualitatively with the help of difference images and SSIM maps for Sparse 16 are shown in Fig. A.4 for T1 in-phase and T2-SPIR, respectively.

Evaluation for regions of interest (ROI)

Further evaluations were performed to compare the performance of the proposed model against the other methods for different regions of interest. For this purpose, the images were segmented into three different regions: liver, kidneys (both left and right), and spleen, with the help of the available segmentation labels from the CHAOS data set with the images. Then the images were cropped to have only the region of interest. Those segmented-cropped images obtained from the results of different methods were

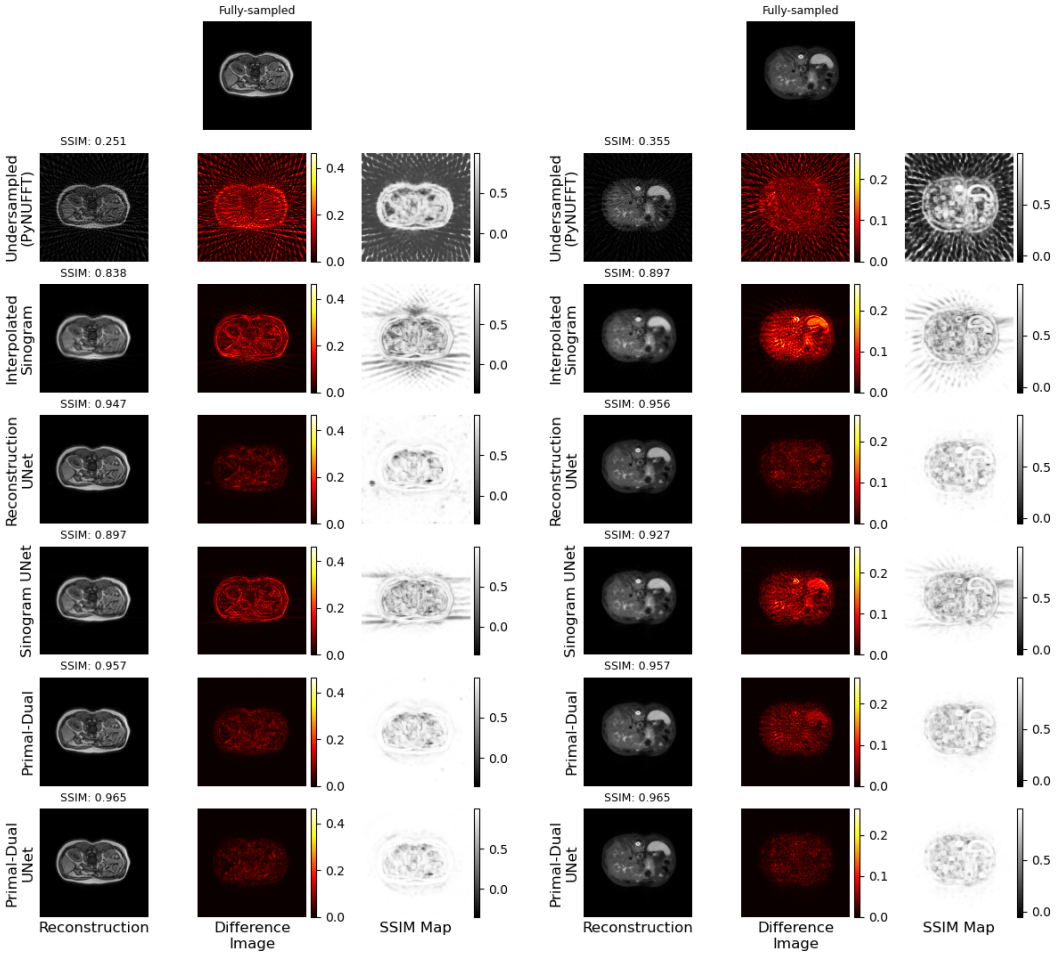


Figure A.4: Qualitative comparisons of the reconstructions of CHAOS Sparse 16 T1w In-phase (left) and T2w (right) (mean±std)

Table A.4: Resultant metrics for liver from CHAOS data set (mean±std)

Method	SSIM	
	Sparse 8	Sparse 16
Undersampled (NUFFT)	0.831±0.083	0.656±0.102
Bilinear Sinogram	0.883±0.046	0.788±0.076
Sinogram UNet	0.893±0.042	0.826±0.063
Reconstruction UNet	0.951±0.029	0.864±0.061
Primal-Dual Network	0.951±0.030	0.866±0.059
Primal-Dual UNet	**0.963±0.026**	**0.888±0.054**
Method	RMSE	
	Sparse 8	Sparse 16
Undersampled (NUFFT)	0.028±0.014	0.065±0.026
Bilinear Sinogram	0.037±0.016	0.048±0.021
Sinogram UNet	0.032±0.014	0.038±0.016
Reconstruction UNet	0.013±0.007	0.026±0.014
Primal-Dual Network	0.012±0.005	0.025±0.011
Primal-Dual UNet	**0.011±0.006**	**0.021±0.010**

then compared against the segmented-cropped version of the ground-truth images. Tables A.5, A.5, A.6 show the quantitative results for liver, kidneys, and spleen. Statistical tests revealed that the proposed model archived statistically significant improvements over all the other methods. Fig. A.5 shows a qualitative comparison of the reconstruction quality for liver, for Sparse 8.

A.3 Discussion

An interesting fact to be noted is that the first-ever 2D MRI was also produced using a back-projection algorithm [LAU73; Gev06]. The only difference between the sparse CT reconstruction and undersampled radial MRI reconstruction using the proposed model (as well as the CT-inspired baselines) is the 1D inverse Fourier transform as preprocessing. The experiments performed as a part of this research show the possibility of using sinogram upsampling techniques combined with FBP to reconstruct undersampled radial MRI.

It was observed that converting the undersampled radial k-space into the corresponding sinogram, then applying bilinear interpolation before finally performing FBP already results in scores better than the undersampled radial MRIs which were reconstructed with the conventional NUFFT. Sinogram UNet, which aims to improve the quality of the bilinearly upsampled sinograms, improves the results even further in terms of SSIM. Even though these two methods perform better than the traditional reconstruction of the

Table A.5: Resultant metrics for both left and right kidneys from CHAOS data set (mean±std)

Method	SSIM	
	Sparse 8	Sparse 16
Undersampled (NUFFT)	0.877±0.180	0.695±0.168
Bilinear Sinogram	0.866±0.083	0.759±0.112
Sinogram UNet	0.881±0.079	0.809±0.095
Reconstruction UNet	0.962±0.054	0.862±0.086
Primal-Dual Network	0.963±0.057	0.866±0.077
Primal-Dual UNet	**0.971±0.049**	**0.892±0.073**

Method	RMSE	
	Sparse 8	Sparse 16
Undersampled (NUFFT)	0.027±0.014	0.070±0.028
Bilinear Sinogram	0.051±0.023	0.065±0.031
Sinogram UNet	0.045±0.020	0.054±0.025
Reconstruction UNet	0.016±0.008	0.035±0.018
Primal-Dual Network	0.016±0.008	0.039±0.023
Primal-Dual UNet	**0.013±0.007**	**0.030±0.014**

Table A.6: Resultant metrics for spleen from CHAOS data set (mean±std)

Method	SSIM	
	Sparse 8	Sparse 16
Undersampled (NUFFT)	0.848±0.149	0.699±0.155
Bilinear Sinogram	0.878±0.114	0.817±0.130
Sinogram UNet	0.885±0.121	0.840±0.127
Reconstruction UNet	0.953±0.104	0.895±0.117
Primal-Dual Network	0.950±0.106	0.894±0.116
Primal-Dual UNet	**0.957±0.118**	**0.907±0.128**

Method	RMSE	
	Sparse 8	Sparse 16
Undersampled (NUFFT)	0.024±0.010	0.059±0.021
Bilinear Sinogram	0.045±0.018	0.055±0.022
Sinogram UNet	0.039±0.017	0.045±0.019
Reconstruction UNet	0.011±0.005	0.023±0.010
Primal-Dual Network	0.012±0.006	0.025±0.013
Primal-Dual UNet	**0.009±0.004**	**0.019±0.009**

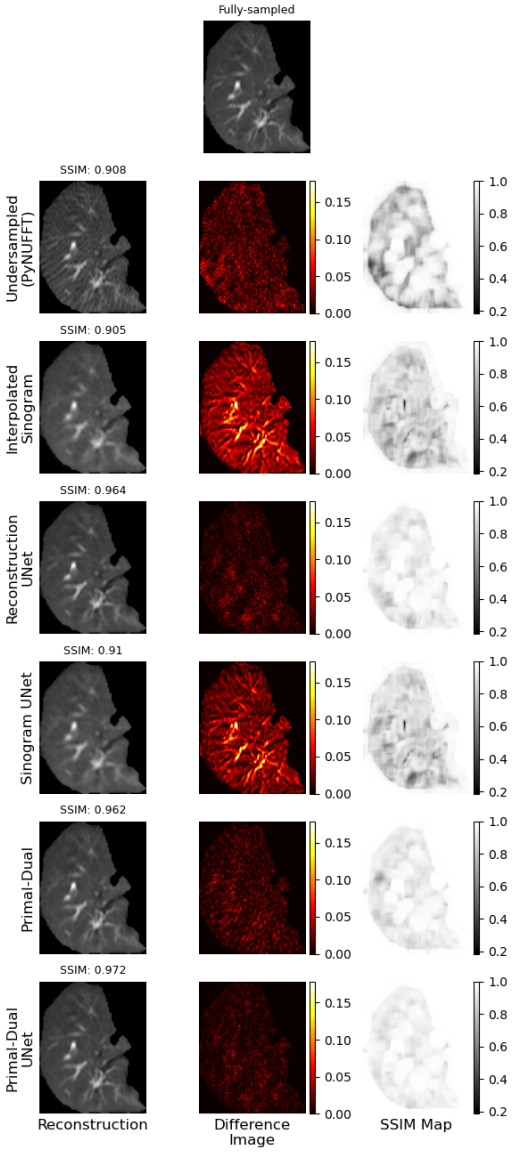


Figure A.5: Qualitative comparisons of the reconstructions of CHAOS T2w Sparse 8: Liver ROI

undersampled radial MRIs using NUFFT, the quantitative and qualitative evaluations yield that they are much inferior to the other three models. It is to be Reconstruction UNet - a well-established model for deep learning based undersampled MRI reconstruction, which is the only model out of the four models which is not CT-inspired and works directly with the reconstructed images, performed very similar to the main baseline model of this paper: Primal-Dual Network. Reconstruction UNet performed better for the brain MRI reconstruction task, whereas the Primal-Dual Network performed better while reconstructing abdominal MRIs. As noted earlier, the proposed method Primal-Dual UNet performed better than all the other methods for all the tasks.

A.4 Primal-Dual UNet for Parallel Beam CT

This section presents the results for the sparse CT reconstruction for the parallel-beam geometry. Each projection consists of 363 detector pixels with a pixel spacing of 1 px from parallel-beams to cover the full axial slice. The sinograms contain 180 equiangular projections with an angular distance of 1° between consecutive projections. Tab. A.7 shows the quantitative comparison of the methods in terms of SSIM (calculated on the normalised intensity values) and RMSE (in the Hounsfield scale), and Fig. A.6, and A.7 portray the range of the resultant SSIM values for the three levels of sparsity: 4, 8, and 16, respectively. In terms of SSIM, the baseline Primal-Dual Network performed better than the proposed method for Sparse 4 and 8. However, the proposed method outperformed the baseline Primal-Dual Network for the highest level of sparsity experimented here: 16. Regarding RMSE, the baseline Primal-Dual Network performed better than the proposed method. However, the proposed method performed better than the other baseline methods - in terms of both, SSIM and RMSE. Finally, qualitative comparisons of the reconstructions with the help of difference images and SSIM maps are shown in Fig. A.8 and A.9, for Sparse 4, 8 and 16, respectively.

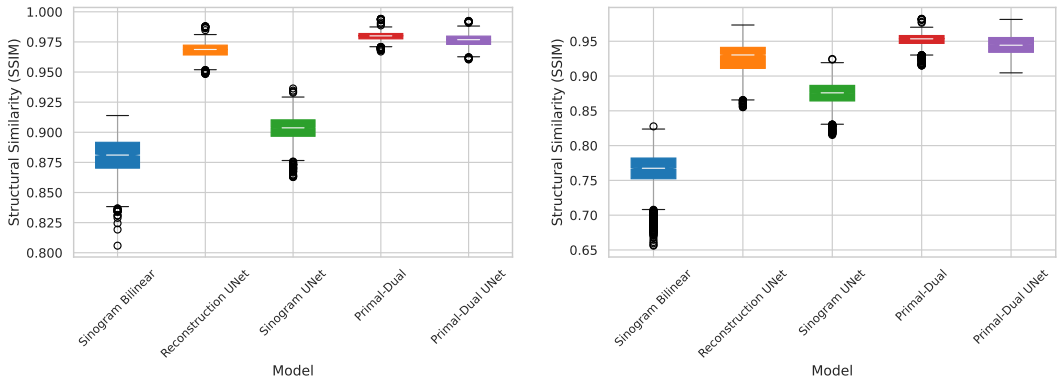


Figure A.6: Box-plots of the resultant SSIM values for CT (parallel-beam geometry) Sparse 4 (left) and Sparse 8 (right)

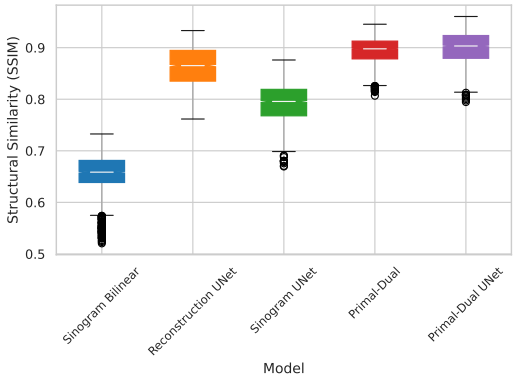


Figure A.7: Box-plots of the resultant SSIM values for CT (parallel-beam geometry) Sparse 16

Table A.7: Resultant metrics for the CT parallel-beam geometry

Method	SSIM		
	Sparse 4	Sparse 8	Sparse 16
Bilinear Sinogram	0.880±0.015	0.764±0.028	0.654±0.038
Sinogram UNet	0.903±0.011	0.875±0.019	0.793±0.037
Reconstruction UNet	0.968±0.006	0.925±0.022	0.863±0.037
Primal-Dual Network	**0.980±0.003**	**0.952±0.010**	0.895±0.026
Primal-Dual UNet	0.976±0.005	0.944±0.014	**0.899±0.032**

Method	RMSE (HU)		
	Sparse 4	Sparse 8	Sparse 16
Bilinear Sinogram	42.504±5.337	72.237±9.285	109.353±13.098
Sinogram UNet	50.024±3.921	53.618±5.887	70.777±11.304
Reconstruction UNet	17.806±2.964	35.180±10.355	62.507±19.468
Primal-Dual Network	**11.913±1.660**	**21.456±4.212**	**43.926±13.337**
Primal-Dual UNet	14.579±2.276	26.545±6.151	46.044±13.362

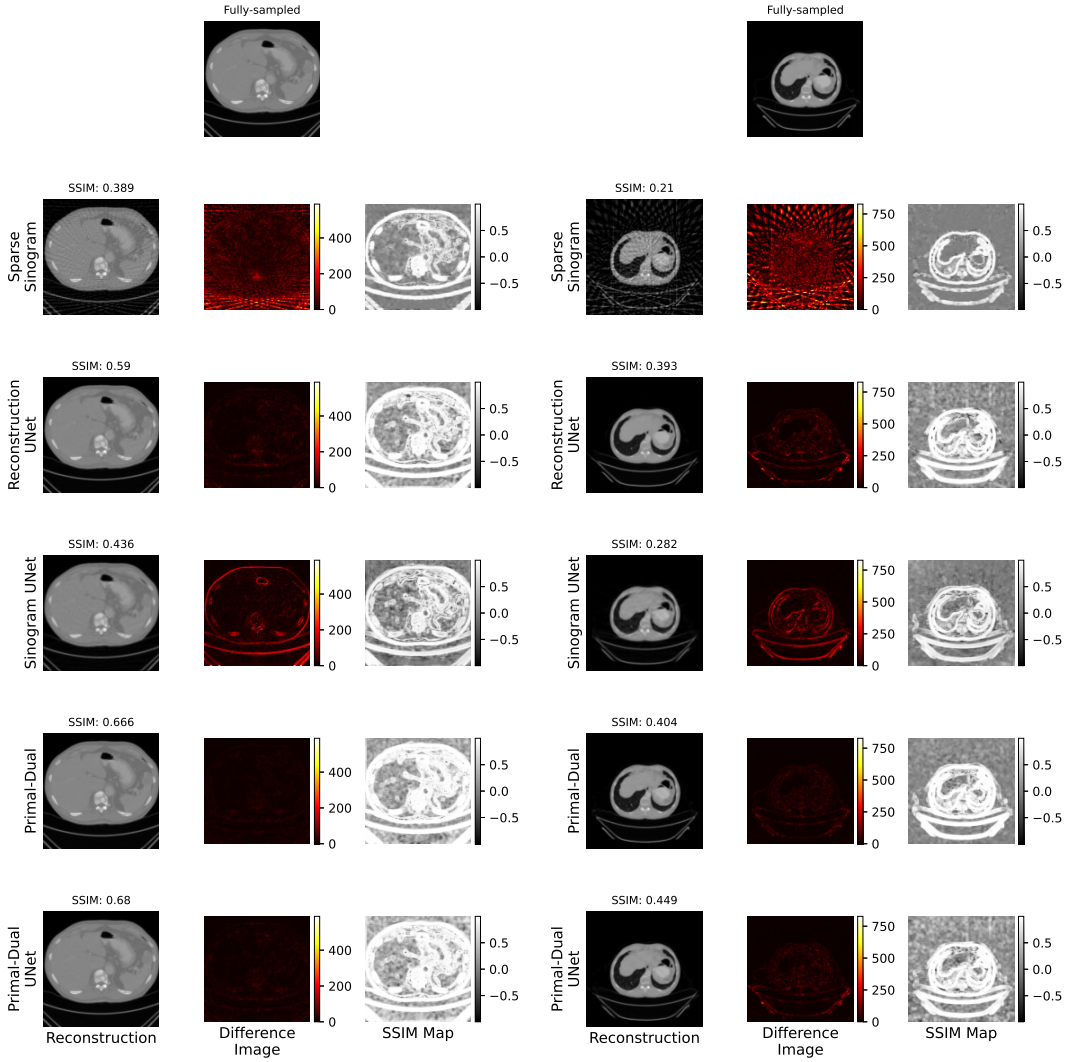


Figure A.8: Qualitative comparisons of the reconstructions of CT (parallel-beam geometry) Sparse 4 (left) and Sparse 8 (right)

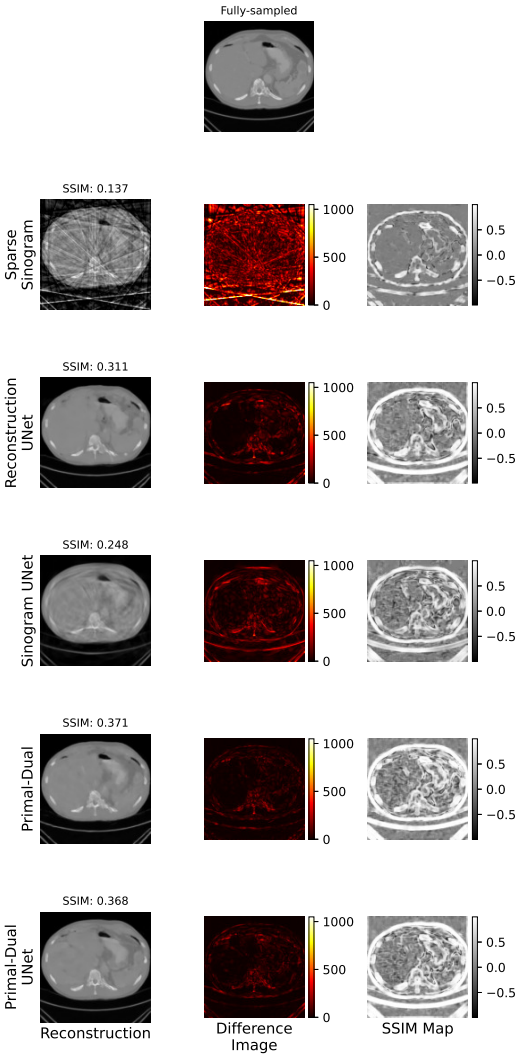


Figure A.9: Qualitative comparisons of the reconstructions of CT (parallel-beam geometry) Sparse 16

Appendix B

Influence of Data Range Parameters in Similarity Metrics

Many related works report errors or similarity using one or several of the metrics described in Sec. 3.8. Some of these widely used metrics need to have set a parameter which describes the range of the values that they operate on. Natural images are usually saved as three-channel 8 bit unsigned integer data, which already sets the range to $[0, 255]$. In some cases, the data might also be available as 32 bit floating point values on the unit interval, such that the range parameter is set to $[0, 1]$.

Unfortunately, pixel values in CT images, despite being directly related to the scanned materials or tissues, merely have a lower bound, i.e. corresponding to no X-ray attenuation. Therefore, there is no consistent value range such that different works likely use different values for these parameters, which negatively affects the comparability. For this reason, the influence of these data range parameters on the error and similarity metrics is shown in the following sections.

B.1 Intensity Scaling in MSE

A useful property of the MSE is about multiplicative scaling of the intensity values. Recall the definition of MSE:

$$\mathrm{MSE}(F, G) = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (f_{ij} - g_{ij})^2. \tag{B.1}$$

Assume that the intensity values are scaled by a positive scalar $s \in \mathbb{R}_{>0}$. Then

$$\begin{aligned}
\mathrm{MSE}(sF, sG) &= \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (s \cdot f_{ij} - s \cdot g_{ij})^2 \\
&= \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (s \cdot (f_{ij} - g_{ij}))^2 \\
&= \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} s^2 \cdot (f_{ij} - g_{ij})^2 \\
&= s^2 \cdot \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (f_{ij} - g_{ij})^2 \\
&= s^2 \cdot \mathrm{MSE}(F, G).
\end{aligned} \tag{B.2}$$

B.2 Normalization in PSNR

Recall the definition of PSNR:

$$\mathrm{PSNR}(F, G) = 20 \log_{10} \left(\frac{I_{max}}{\mathrm{MSE}(F, G)} \right), \tag{B.3}$$

where I_{max} is the maximum intensity value, i.e. the normalization value. Let us assume that it was set to an incorrect value $I_{ic} \in \mathbb{R}_{>0}$, resulting in the incorrect PSNR function

$$\mathrm{PSNR}_{ic}(F, G) = 20 \log_{10} \left(\frac{I_{ic}}{\mathrm{MSE}(F, G)} \right). \tag{B.4}$$

The goal is now to relate PSNR and PSNR_{ic}. Since the normalization values are scalars, it is possible to find a value $\alpha \in \mathbb{R}_{>0}$ such that $I_{max} = \alpha I_{ic}$. Putting this into Eq. B.3 results in

$$\begin{aligned}
\mathrm{PSNR}(F, G) &= 20 \log_{10} \left(\frac{I_{max}}{\mathrm{MSE}(F, G)} \right) \\
&= 20 \log_{10} \left(\frac{\alpha I_{ic}}{\mathrm{MSE}(F, G)} \right) \\
&= 20 \left(\log_{10}(\alpha) + \log_{10} \left(\frac{I_{ic}}{\mathrm{MSE}(F, G)} \right) \right) \\
&= \underbrace{20 \log_{10}(\alpha)}_{=:\tilde{\alpha}} + \underbrace{20 \log_{10} \left(\frac{I_{ic}}{\mathrm{MSE}(F, G)} \right)}_{\mathrm{PSNR}_{ic}(F,G)} \\
&= \mathrm{PSNR}_{ic}(F, G) + \tilde{\alpha}.
\end{aligned} \tag{B.5}$$

Thus, the correct PSNR value can be calculated by simply adding a constant to the incorrect PSNR value. This shows that the normalization factor has merely an additive influence on the PSNR calculation. Consequently, relations between two PSNR values remain unchanged independent of the normalization factor and, moreover, even differences between two PSNR values are not influenced by this factor.

Not only do these properties hold for the normalization factor I_{max} but also for multiplicative scalings of the actual intensity values of F and G. Using the property of the previous section, $\mathrm{MSE}(sF, sG) = s^2 \cdot \mathrm{MSE}(F, G)$ (Eq. B.1), it follows

$$\begin{aligned}
\mathrm{PSNR}(sF, sG) &= 20 \log_{10} \left(\frac{I_{max}}{\mathrm{MSE}(sF, sG)} \right) \\
&= 20 \log_{10} \left(\frac{I_{max}}{s^2 \mathrm{MSE}(F, G)} \right) \\
&= 20 \log_{10} \left(\frac{\overbrace{s^{-2} I_{max}}^{=:\alpha I_{ic}}}{\mathrm{MSE}(F, G)} \right) \\
&\overset{\text{(Eq. B.5)}}{=} \mathrm{PSNR}(F, G) + \tilde{\alpha}.
\end{aligned} \tag{B.6}$$

B.3 Dynamic Range in SSIM

Recall the definition of SSIM [WBS+04]:

$$\begin{aligned}
\mathrm{SSIM}(x, y) &= \frac{(\overbrace{2\mu_x\mu_y}^{=:a_1} + k_1^2 L^2)(\overbrace{2\sigma_{xy}}^{=:a_2} + k_2^2 L^2)}{(\underbrace{\mu_x^2 + \mu_y^2}_{=:a_3} + k_1^2 L^2)(\underbrace{\sigma_x^2 + \sigma_y^2}_{=:a_4} + k_2^2 L^2)} \\
&= \frac{(a_1 + k_1^2 L^2)(a_2 + k_2^2 L^2)}{(a_3 + k_1^2 L^2)(a_4 + k_2^2 L^2)} \\
&= \frac{(\frac{a_1}{k_1^2} + L^2)(\frac{a_2}{k_2^2} + L^2)}{(\frac{a_3}{k_1^2} + L^2)(\frac{a_4}{k_2^2} + L^2)} \\
&= \frac{(b_1 + L^2)(b_2 + L^2)}{(b_3 + L^2)(b_4 + L^2)} \\
b_1 := \frac{a_1}{k_1^2},\ b_2 &:= \frac{a_2}{k_2^2},\ b_3 := \frac{a_3}{k_1^2},\ b_4 := \frac{a_4}{k_2^2}
\end{aligned} \tag{B.7}$$

As in the previous section, assume that the dynamic range variable was chosen incorrectly, such that $L = \alpha L_{ic}$.

$$\begin{aligned}\mathrm{SSIM}(x,y) &= \frac{(b_1+\alpha^2L_{ic}^2)(b_2+\alpha^2L_{ic}^2)}{(b_3+\alpha^2L_{ic}^2)(b_4+\alpha^2L_{ic}^2)} \\ &= \frac{(b_1+c)(b_2+c)}{(b_3+c)(b_4+c)},\ c := \alpha^2L_{ic}^2 \\ &= \frac{c^2+(b_1+b_2)c+b_1b_2}{c^2+(b_3+b_4)c+b_3b_4}\end{aligned} \tag{B.8}$$

This is a quotient of two quadratic functions (wrt. the variable c) which can be converted to the form [Gre55]:

$$\begin{aligned}y &= \frac{\lambda(px+q)^2+\mu(rx+s)^2}{\lambda'(px+q)^2+\mu'(rx+s)^2} \\ &= \frac{\lambda(p^2x^2+2pqx+q^2)+\mu(r^2x^2+2rsx+s^2)}{\lambda'(p^2x^2+2pqx+q^2)+\mu'(r^2x^2+2rsx+s^2)} \\ &= \frac{(\lambda p^2+\mu r^2)x^2+2(\lambda pq+\mu rs)x+(\lambda q^2+\mu s^2)}{(\lambda' p^2+\mu' r^2)x^2+2(\lambda' pq+\mu' rs)x+(\lambda' q^2+\mu' s^2)}\end{aligned} \tag{B.9}$$

Depending on the values of the variables p, q, r, s, λ, λ', μ and μ', the graph can

- lie entirely on one side of an asymptote,
- have exactly one turning point,
- have exactly two turning points.

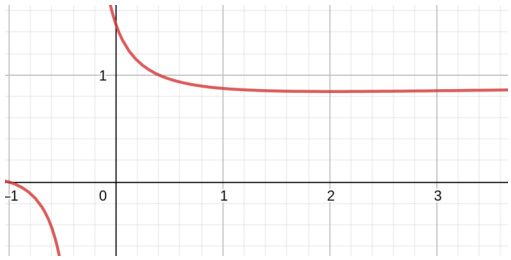
Since the derivation of the sought variables in Eq. B.9 from the fraction in Eq. B.8 is not straightforward, other properties of the variables in the equation can be exploited to facilitate the evaluation. b_1 through b_4 are derived from the mean values and the standard deviations, scaled by the positive constants k_1 or k_2. Assuming images with non-negative values, both mean values and standard deviations are non-negative, as well. Therefore, b_1 through b_4 are non-negative. Moreover, the SSIM is symmetric in b_1 and b_2 as well as in b_3 and b_4. The definition of $c := \alpha^2L_{ic}^2$ implies $c \geq 0$ (in fact, $c > 0$ since both α and L are positive to be semantically correct).

Because of these properties, it holds $\lim_{c\to\infty} \mathrm{SSIM} = 1$, $\forall c : \mathrm{SSIM} \geq 0$, and it is simple to find the cases where, in fact, $\mathrm{SSIM} \leq 1$:

$$\begin{aligned}&\frac{c^2+(b_1+b_2)c+b_1b_2}{c^2+(b_3+b_4)c+b_3b_4} \leq 1 \\ &c^2+(b_1+b_2)c+b_1b_2 \leq c^2+(b_3+b_4)c+b_3b_4 \\ &\underbrace{(b_1+b_2-b_3-b_4)}_{=:S}\,c \leq -\underbrace{(b_1b_2-b_3b_4)}_{=:P} \\ &\Leftrightarrow \begin{cases} c \leq -\frac{P}{S}, & \text{if } S > 0 \\ c \geq -\frac{P}{S}, & \text{otherwise} \end{cases}\end{aligned} \tag{B.10}$$

(a) P $<$ 0, S $<$ 0

(b) P $>$ 0, S $<$ 0



(c) P $<$ 0, S $>$ 0

(d) P $>$ 0, S $>$ 0

Figure B.1: Graphs of exemplary SSIM functions dependent on the dynamic range variable c for the four different cases derived in the text (SSIM value on ordinate, c on abscissa).

One can now evaluate the four different cases where SSIM ≤ 1 (exemplary graphs in Fig. B.1):

	$P \leq 0$	$P > 0$
$S < 0$	$\forall c : c \geq -\frac{P}{S}$	$\exists c : c \geq -\frac{P}{S}$
$S > 0$	$\exists c : c \leq -\frac{P}{S}$	$\nexists c : c \leq -\frac{P}{S}$

Setting the partial derivative of SSIM wrt. c to zero results in the solutions

$$\begin{aligned} c_E &= -\frac{2P \pm \sqrt{4P^2 - 4S(b_1b_2b_3 + b_1b_2b_4 - b_1b_3b_4 - b_2b_3b_4)}}{2S} \\ &= -\frac{2P \pm \sqrt{4P^2 - 4S(b_1b_2(b_3 + b_4) - b_3b_4(b_1 + b_2))}}{2S} \end{aligned} \tag{B.11}$$

providing possible candidates for extreme points.

Case 1: $P \leq 0 \wedge S < 0$**.** There is either no non-negative extreme point or exactly one at c_E^+. If there is none, SSIM is increasing wrt. c. Otherwise, it is decreasing in $[0, c_E^+)$ and increasing in $[c_E^+, \infty)$, and $\lim_{c\to\infty} \text{SSIM} = 1^-$.

Case 2: $P > 0 \wedge S < 0$**.** Assuming a lower bound c^* was found s.t. SSIM < 1, there is exactly one non-negative extreme point c_E^+ s.t. SSIM is decreasing in $[0, c_E^+)$ and increasing in $[c_E^+, \infty)$, and $\lim_{c\to\infty} \text{SSIM} = 1^-$.

Case 3: $P \leq 0 \wedge S > 0$**.** Assuming an upper bound c^* was found s.t. SSIM < 1, there is exactly one non-negative extreme point at c_E^+ s.t. SSIM is increasing in $[0, c_E^+)$ and decreasing in $[c_E^+, \infty)$. Since c^* is an upper bound not to invalidate SSIM ≤ 1, it holds $c^* < c_E^+$ and therefore, SSIM is increasing in $[0, c^*]$, and $\lim_{c\to\infty} \text{SSIM} = 1^+$.

Case 4: $P > 0 \wedge S > 0$**.** No c fulfills the requirement SSIM ≤ 1.

Unfortunately, there is no simple calculation to rectify SSIM values when an incorrect dynamic range was used. Attempting to modify the dynamic range variable incorporates knowledge about statistics of the images that are compared, which might not be available anymore. Aggravatingly, SSIM is not injective wrt. the dynamic range variable in some cases, in particular for $P > 0 \wedge S < 0$, such that different values of the dynamic range can result in the same SSIM values.

A rather interesting insight into the SSIM is Case 4: If both P and S are positive, the SSIM is always larger than one, independently of the choice of the dynamic range variable. As an example, take $b_1 = 1$, $b_2 = 4$, $b_3 = 1.5$, and $b_4 = 2$. Tracing these values back to the variables in the original Eq. B.7 (assuming $k_1 = 0.01$ and $k_2 = 0.03$ as the original paper, and, without loss of generality, $\alpha = 1$) leads to

$$\begin{aligned} \mu_x\mu_y &= 5 \cdot 10^{-5}, & \sigma_{xy} &= 0.0018 \\ \mu_x^2 + \mu_y^2 &= 15 \cdot 10^{-5}, & \sigma_x^2 + \sigma_y^2 &= 0.0018 \end{aligned}$$

which might seem unrealistically small for natural images with values in $[0, 255]$, but it can easily result from images made up of X-ray attenuation coefficients with non-negative values with the 99th percentile being ≈ 0.03.

Appendix C

Medical Evaluation Tool

This final chapter focuses on the evaluation of the image quality of reconstruction algorithms for actual medical use by doctors.

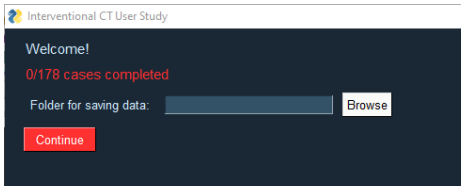
As previously stated in Sec. 3.8, the metrics for assessing the quality of reconstructed images or volumes do not necessarily resemble the quality that is desired by doctors to guide them visually in certain use cases. For this task-based assessment, a simple medical evaluation tool was developed in the course of this thesis.

C.1 Requirements and Work Flow

The tool needs to meet some requirements to make the evaluation process as simple and least time-consuming for the user, i.e. doctors, while providing as much information about single samples as well as unbiased statistical properties of the entire data set as possible. Consequently, the following work flow had to be established:

- Setting a folder for saving the results of the processed samples (Fig. C.1 (a)).
- Ability to continue the assessment with the remaining unprocessed samples after closing the tool.
- The data set should be processed randomly to avoid biases in the evaluation process.
- A progress bar shows how many samples have been processed already.
- For each sample, the predictions of multiple models are visualized clearly next to the ground truth in a first screen (Fig. C.1 (b)):
  - The user needs to rank each prediction.
  - Hovering over a sample creates a cross-hair in all shown samples at the same location. Right-clicking on the sample shows the Hounsfield Units of the sample at the position of the cross-hair for all shown samples.

(a)



(b)

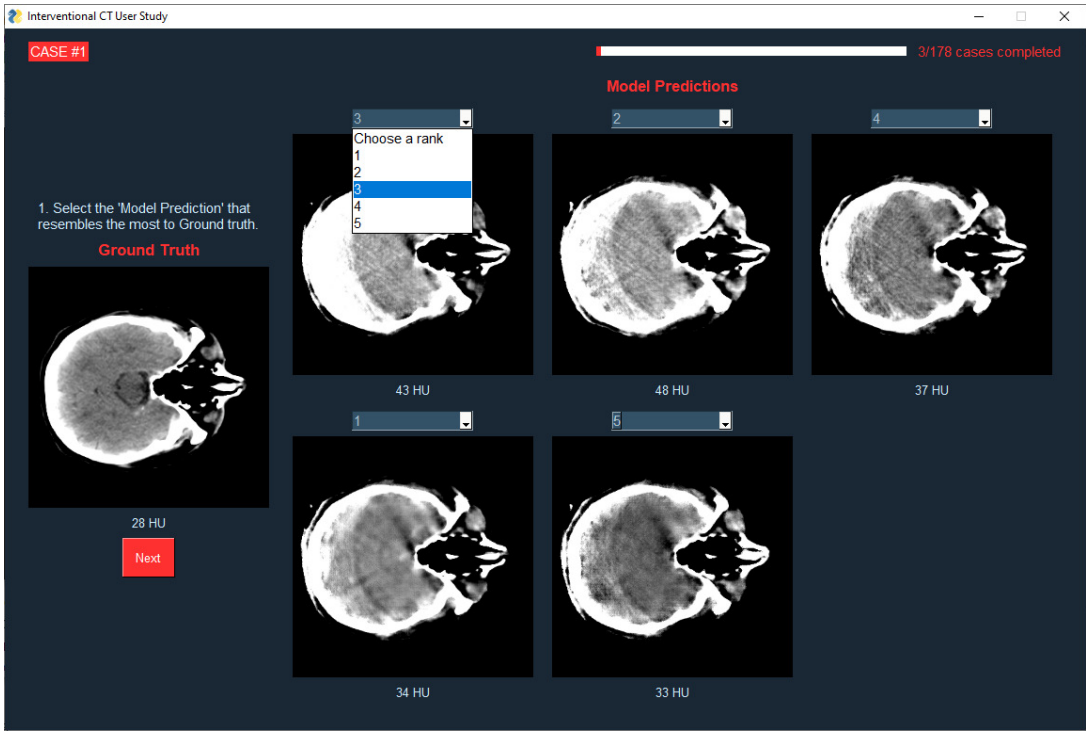
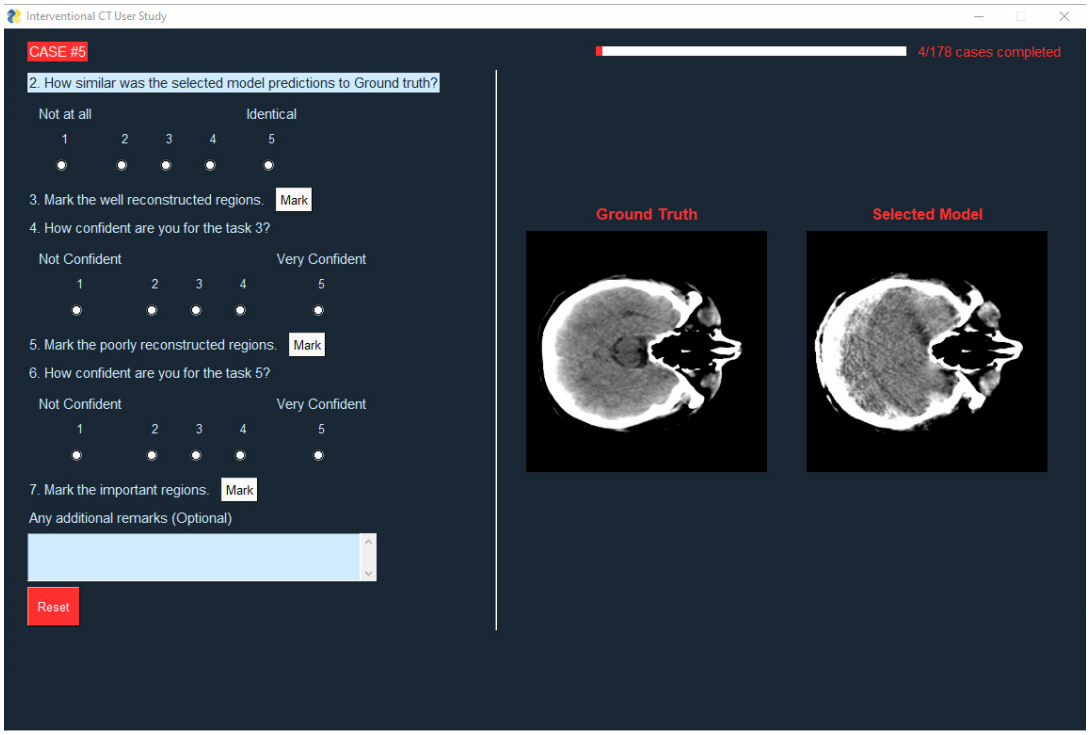


Figure C.1: First stages of the evaluation tool. (a) Landing page with the possibility to select a folder where the annotations are saved to. (b) Stage 1 for an exemplary slice. Each output of randomized models is given a rank. Right-clicking at an arbitrary position inside one of the images reveals the Hounsfield Units for all images at the same position below the images.

(a)



(b)

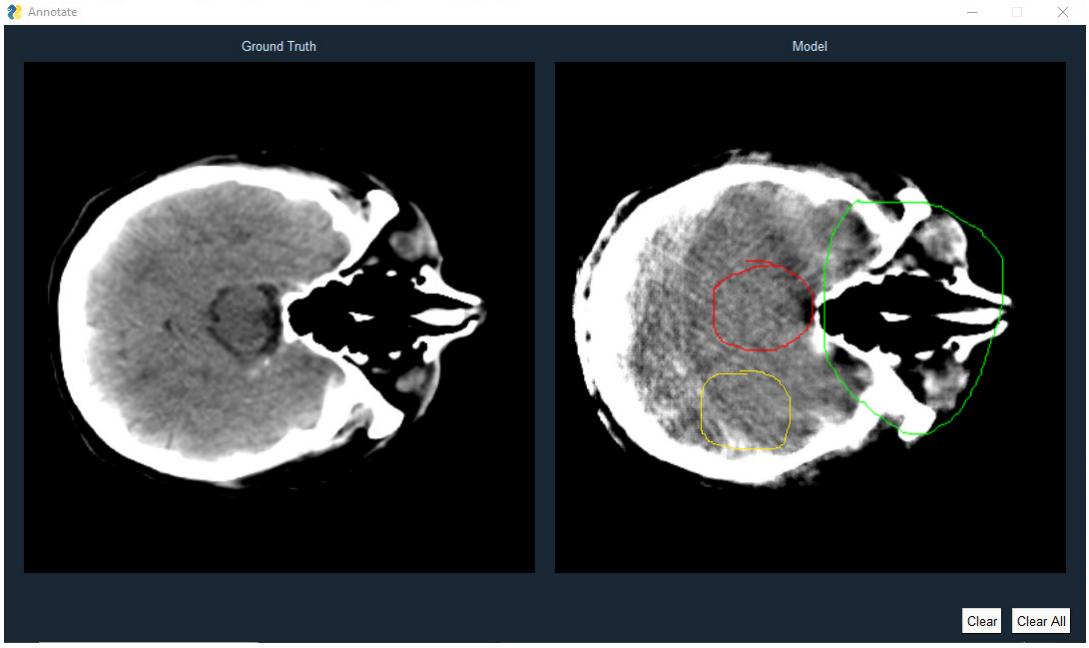


Figure C.2: Second stage of the evaluation tool. (a) Answering the questions on the left while comparing the best ranked image to the ground truth. (b) Marking well (green) and poorly reconstructed (green) regions, as well as important (yellow) regions.

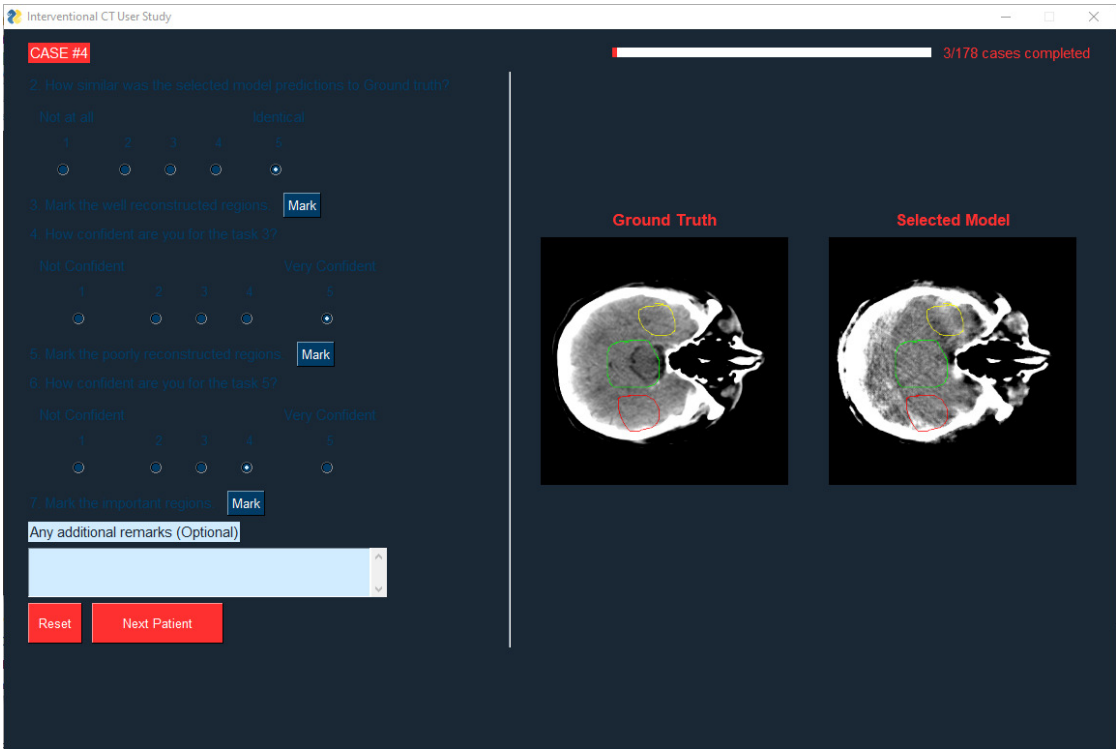


Figure C.3: Exemplary fully completed second stage of one slice. If necessary, the user can reset the annotations of the current slice and restart from Stage 1.

- After ranking the predictions of the models, the user continues to a second screen for a more detailed assessment of the model prediction with the chosen highest rank (Fig. C.2 (a)):
  - The user determines how similar the prediction with the chosen highest rank resembles the ground truth.
  - The user needs to mark well and poorly reconstructed as well as important regions in the prediction (Fig. C.2 (b)).
  - For each selection, the user is asked how confident his decision was.
  - The user is given the possibility to add additional remarks as free text.
  - The user can reset the assessment of the current sample.

When the user has finished annotating a sample, the screen looks similar to Fig. C.3 before continuing to Stage 1 of the next sample.

C.2 Implementation Details

The tool is intended to run platform-independently. For this reason, Python 3.9.7 is used in conjunction with PySimpleGUI 4.56, and to enable quick prototyping with the ability to refine certain aspects by having access to the low-level backend functions (in this case: Tk via the Python interface Tkinter). The graphical user interface is event-driven and the program is designed in an object-oriented way:

- `class UserStudy`: The entry point of the program. It is responsible for setting up the application with its UI elements and event handlers.
- `class SliceAnnotation`: Defines methods for setting up a window for annotating slices with well or poorly reconstructed or important regions.
- `class Ui_UserStudy` and `class Ui_SliceAnnotation`: Factory classes to generate the UI layouts of `class UserStudy` and `class SliceAnnotation`.
- `class PatientLoader`: This class is responsible for loading the samples (both model predictions and ground truth) and keeping track of the index to enable processing all samples in the data set. Moreover, the samples are loaded asynchronously to avoid freezing of the application while loading.
- `class UserInput`: A data class containing the selections and choices of the user for one sample.

Though platform-independent by using Python, users should not be expected to be able to create a Python virtual environment including all the necessary packages and start the program from a command-line interface. For this reason, PyInstaller 4.7 was used to pack all necessary dependencies of the program and compile an executable file. All these files are compressed into a zip archive, which makes the distribution of the program very simple, as it can also already contain the data set to be assessed.

The tool can be downloaded from Github[1].

[1] `https://github.com/suhitaghosh10/medical-evaluation-tool`