Dissertation zur Erlangung des Grades des Doktors der Ingenieurwissenschaften der Naturwissenschaftlich-Technischen Fakultäten der Universität des Saarlandes

Tomographic Reconstruction of Combined Tilt- and Focal Series in Scanning Transmission Electron Microscopy

Tim Dahmen
Saarbrücken, 2015

UNIVERSITÄT
DES
SAARLANDES

Datum des Kolloquiums	24. September 2015
Dekan der Fakultät 6	Prof. Dr. Markus Bläser
Betreuer der Promotion	Prof. Dr.-Ing. Philipp Slusallek
Prüfungsausschuss	Prof. Dr. Thorsten Herfet (Vorsitzender) Prof. Dr.-Ing. Philipp Slusallek Prof. Dr. rer. nat. Dr. h. c. mult. Alfred K. Louis Prof. Dr. Wolfgang Heidrich Dr.-Ing. Richard Membarth (akademischer Beisitzer)

Contents

Zusammenfassung **6**

Abstract **8**

1 Introduction **10**

1.1 Background and Motivation 10

1.2 Related Work 14

1.2.1 Tilt Series Tomography 14

1.2.2 Focal Series Tomography 17

1.2.3 Exotic Sources of Three Dimensional Resolution 17

1.2.4 Reconstruction Methods 18

1.3 Own Contributions 21

1.4 Thesis Structure 22

2 Theory **24**

2.1 Definition of the Term “Resolution” 24

2.1.1 Resolution Limit by the Rayleigh Criterion 24

2.1.2 Resolution Limit by Nyquist-Shannon Theorem 26

2.1.3 The Full-Width-at-Half-Maximum Criterion 26

2.1.4 Contrast and Electron Statistics 27

2.1.5 Resolution in Three Dimensions: Isotropy and Axial Elongation 29

2.2 The Ray Transform for Parallel Illumination 29

2.3 Tomographic Reconstruction Methods 30

2.4 The STEM Forward Projection 34

2.4.1 Double Cone Model of the Electron Beam 34

2.4.2 The STEM Transform as a Linear Operator 35

2.4.3 Relation to the Parallel Ray Transform 38

2.4.4 The Fourier Transform of the STEM Transform 39

2.4.5 The Adjoint of the STEM Transform 41

3 Implementation and Evaluation 44

3.1 Alignment 44

3.1.1 Lateral Alignment 45

3.1.2 Axial Alignment 45

3.1.3 Particle Chain Detection 46

3.1.4 Nanoparticle Position Triangulation 46

3.1.5 Estimation of Focal Values 47

3.2 Tomographic Reconstruction 47

3.2.1 Forward Projection 48

3.2.2 Unmatched Back Projection Based on Heuristic Weighting Factor 52

3.2.3 Matched Back Projection Based on Adjoint Transform . 54

3.2.4 Update Loop 56

3.3 Evaluation and Results 57

3.3.1 Sample Preparation 58

3.3.2 Data Acquisition 58

3.3.3 Evaluation of Alignment 59

3.3.4 Evaluation of Unmatched Back Projection 60

3.3.5 Evaluation of Matched Back Projection 64

3.3.6 Results on Experimental Data 66

3.3.7 Results on Phantom Data 67

3.4 Discussion of the Reconstruction Algorithm 69

3.4.1 Computational Cost per Iteration 69

3.4.2 Classification of the TF-ART Algorithm 70

3.4.3 Influence of the Matched Back Projection 71

3.4.4 Conclusions of Experimental Evaluation 71

4 Software Architecture 73

4.1 Existing Software Packages 73

4.1.1 Open Source Packages 73

4.1.2 Commercial Packages 75

4.2 Requirements: Extensibility, Modularity, and Performance . . . 76

4.3 The Ettention Architecture 77

4.3.1 Overview 77

4.3.2 Plug-in Concept and Binaries from Kernel Files 79

4.3.3 Memory Management 79

4.3.4 Parallelism 80

4.4 Individual Building Blocks 81

4.5 Interfacing with Other Software 85

4.6 Results 86

4.6.1 Evaluation Methods 86

4.6.2 Reconstruction Quality 86

4.6.3 Performance 88

4.7 Discussion of the Software Architecture 91

5 Conclusions 94

6 Future Work 97

6.1 Applicability and Electron Dose Restrictions 97

6.2 Regularization and A-priori Information 98

6.3 Towards Beam Blurring Correction 98

References **103**

List of Abbreviations **116**

List of Symbols **118**

Appendix: Derivation of the Relationship Between the STEM Transform and the Parallel Ray Transform **119**

Zusammenfassung

In dieser Arbeit wird Tomographie unter Nutzung von Rastertransmissionselektronenmikroskopen (STEM) mit begrenzter Tiefenschärfe untersucht. Wir schlagen ein neues Aufnahmeschema für die High-Angle Annular Dark-Field (HAADF) Tomographie vor: Die kombinierte Kipp- und Fokusserie (CTFS). In diesem Schema wird drei-dimensionale (3D) Information gewonnen, indem eine Probe mechanisch gedreht wird. Für jede Richtung wird eine Serie von Bildern mit unterschiedlicher Fokustiefe aufgenommen.

Wir stellen die STEM-Transformation vor, eine Verallgemeinerung der bekannten Strahl-Transformation (Röntgen-Transformation) für parallele Strahlen. Die STEM-Transformation berücksichtigt die konvergente Form des Elektronenstrahls in aberrationskorrigierten STEM. Die Abbildung wird analytisch untersucht und es wird gezeigt, dass es sich (1) um eine lineare Faltung handelt sowie (2) um einen selbstadjungierten Operator.

Wir stellen einen iterativens Algorithmus für die tomographischen Rekonstruktion von CTFS Daten vor. Der Algorithmus nutzt die Kaczmarz Methode um das System $Ax = b$ approximativ im Sinne der kleinsten Quadrate zu lösen. Hierbei sind b die Bilder, x das gesuchte Tomogramm und A die sogenannte Systemmatrix. Da A für eine explizite, selbst dünn besetzte, Speicherung zu groß ist, wird die Matrix implizit ausgedrückt. Dies geschieht durch eine Vorwärts- und eine Rückprojektion. Die Vorwärtsprojektion nutzt eine Implementierung der STEM-Transformation auf Basis von stochastischem Raytracing. Es werden zwei unterschiedliche Rückprojektionen definiert, "paarweise" und "nicht paarweise". Die nicht paarweise Rückprojektion nutzt einen heuristischen Gewichtungsfaktor, wohingegen die paarweise Version auf einer numerischen Approximation der adjungierten STEM-Transformation basiert. Diese Implementierung nutzt im Fourier-Raum vorberechnete Faltungsoperationen und lineare Interpolation.

Eine experimentelle Evaluierung des Algorithmus zeigt, dass die kombinierte Kipp- und Fokusserie die Artefakte "axiale Verlängerung" sig-

nifikant reduziert, d.h. sie führt bei gleichem Drehbereich zu einer isotroperen Auflösung als Verfahren auf Basis einer reinen Kipp- oder reinen Fokusserie. Die paarweise Rückprojektion weist ein drastisch schnelleres Konvergenzverhalten auf als die nicht paarweise Version.

Zum Abschluss wird das Softwarepaket "Ettention" präsentiert, das eine große Bandbreite tomographischer Rekonstruktionsprobleme mittels iterativer Verfahren löst. Es wird gezeigt, wie die scheinbar widersprüchlichen Anforderungen "Erweiterbarkeit", "Modularität" und "Performanz" gleichzeitig erfüllt werden können, indem ein Werkzeugkasten mit Bausteinen für iterative Rekonstruktionsverfahren zur Verfügung gestellt wird. Diese Bausteine können schnell zu applikationsspezifischen Rekonstruktionsalgorithmen kombiniert werden.

Abstract

In this thesis, we investigate tomography using scanning transmission electron microscopy (STEM) with limited depth of field. A combined tilt- and focal series (CTFS) is proposed as a new recording scheme for high-angle annular dark-field STEM tomography. Hereby, three-dimensional (3D) data is acquired by mechanically tilting the specimen and at each tilt direction recording a series of images with different focal depth (focal series).

The STEM transform is introduced as a generalization of the well-known ray transform for parallel illumination that considers the convergent shape of an electron beam in aberration corrected STEM. The STEM transform is investigated analytically and it is shown that it is (1) a linear convolution and (2) self-adjoint.

We introduce an iterative algorithm to solve the problem of tomographic reconstruction for data recorded with this new scheme. The algorithm uses the Kaczmarz method to solve the system $Ax = b$ in a least-squares sense, where b is the data, x the searched-for tomogram, and A is the so-called system matrix. As the system matrix is too large for an explicit, even sparse, representation, A is expressed implicitly by means of a forward and a back projection. The forward projection used in this thesis is an implementation of the STEM transform based in stochastic ray-tracing. Two different back projections are proposed. The first is based on a heuristic weighting factor and called “unmatched”. The second method is based on an numeric approximation of the adjoint STEM transform and called “matched”. The implementation uses precomputed convolution operations and linear interpolation to achieve high computational efficiency.

By experimental evaluation of the algorithm we show that the method significantly reduces the artifacts known as axial elongation, i.e. leads to a more isotropic resolution than pure tilt series based approaches at the same tilt range as well as pure focal series methods. Furthermore, the matched back projection converges drastically faster than the unmatched version.

Finally, the “Ettention” software package is introduced as a platform to implement a wide range of tomographic reconstruction problems. It is shown how the seemingly contradictory requirements “extensibility”, “modularity” and “performance” can be achieved at the same time by providing a toolbox of building blocks, which can quickly be assembled to application specific reconstruction algorithms.

Chapter 1

Introduction

1.1 Background and Motivation

Tomographic reconstruction, sometimes called “image reconstruction from projections”, is the problem to reconstruct an image function from transmission data obtained by illuminating an object with radiation from many directions. The problem is of enormous practical relevance and has numerous applications in fields such as computed tomography in medical imaging and material science, three-dimensional (3D) electron microscopy, meteorology and geology to name but a few. The different applications require the use of different types of radiation. Images have been recorded with X-Ray radiation of different spectra, electron radiation, and even optical light.

Of particular interest for this thesis is the imaging using scanning transmission electron microscopy (STEM) with a high angle annular dark field (HAADF) detector. In this microscopy mode, a focused electron beam is scanned over the specimen and the image is formed sequentially, pixel-by-pixel. The STEM imaging mode is a transmissive mode, i.e. the detector is on the opposite side of the specimen as seen from the electron source such that radiation has to pass through the specimen in order to generate a signal. The detector consists of an annular ring with a high inner and outer cutoff radius. This way, only those electrons contribute to the detector signal that are scattered more than a minimal scattering angle. This setup gives rise to atomic number contrast (Z-contrast), i.e. the intensity of the signal is a function of the atomic numbers of the sample volume that interacted with the beam.

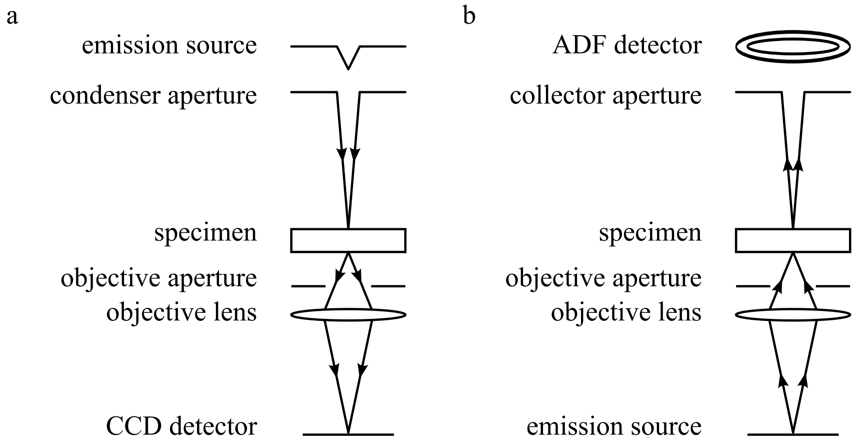

Figure 1.1: Ray diagrams for imaging in a) the conventional transmission electron microscope (CTEM) with charge-coupled device (CCD) detector and b) the scanning transmission electron microscope (STEM) with annular dark field (ADF) detector.

Because there is no objective lens between the specimen and the detector (Figure 1.1), the imaging mode is particularly suitable for imaging thick specimen, i.e. samples with a thickness of 1 μm and higher. On the other hand, because of the nature of Z-contrast, the specimen must either consist of or be stained with heavy atoms, because light elements give insufficient contrast.

As with all electron microscopes, the primary output of HAADF-STEM imaging is a two-dimensional (2D) image of the specimen, i.e. one projection. In order to retrieve the 3D shape of the object, computational methods are applied to combine multiple 2D images into one 3D volume. Several approaches exist as explained below but the most commonly used method is tilt series based tomography. Hereby, the specimen is rotated and images are taken from different directions. Tomographic reconstruction algorithms are used to compute a 3D volume from the projections. Interestingly, this operation is the inverse operation of volume rendering, where a 2D image is generated from a volumetric object. Therefore, one possible view is to consider tomography as an inverse rendering problem.

Either way, the tilt range strongly influences the resolution of the 3D reconstructions. Therefore, one would ideally acquire tilted images covering the entire angular range of ±90°. However, in practice, the maximum tilt range is usually only about ±60-78° due to mechanical limitations of specimen holders and because the effective thickness of the specimen as seen by the electron beam increases as the section is tilted (Figure 1.2). The tomographic reconstruction then suffers from missing information and limited resolution on account of

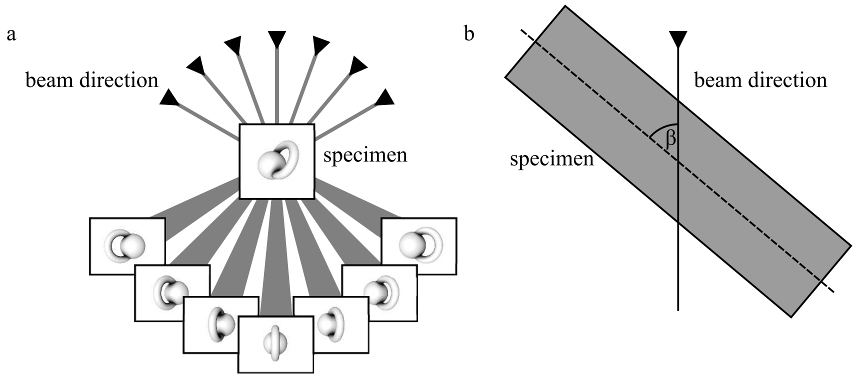

Figure 1.2: a) The principle of tomographic reconstruction. A specimen is rotated in a transmission microscope and projections are taken from several directions. The projections are then computationally combined to reconstruct the volume. b) Connection between the missing wedge problem and the specimen thickness. As the specimen is tilted by an angle β, the thickness "as seen by the electron beam" increases by $1/\cos\beta$. Therefore, some projection directions are not available for the tomographic reconstruction.

this "incomplete projection set". Depending of the used recording scheme, the missing region has the shape of a "missing wedge" (single-tilt), "missing pyramid" (double tilt) or "missing cone" (conical tilt). Independent of the used recording scheme, tilt series based methods are typically recorded with the best parallel illumination possible to reach a high depth of field (DOF) and avoid blurring of the image.

Some techniques avoid tilting altogether. In aberration-corrected STEM, the electron beam is convergent with a focal depth of typically a few nanometers. This is a side-effect of the spherical aberration correctors and required to reach the lowest possible spot size of the electron beam. At the same time, the convergent electron beam leads to images with a limited DOF. This allows a completely different approach to tomography. A focal series can be used to obtain 3D STEM information. Hereby, axial information is retrieved from a stack of images with different focal values and the 3D data set is deconvolved to remove the blurring artifacts resulting of the limited DOF.

In both cases, tilt- and focal series tomography, the resolution suffers from anisotropic artifacts, called "axial elongation" (Figure 1.3). This means that the resolution is lower in axial direction than in lateral direction and objects appear blurred in beam direction. Because the blurring severely hinders the interpretation of 3D images, axial elongation is considered one of two main limiting factors of 3D electron microscopy. The other limiting factor is the low

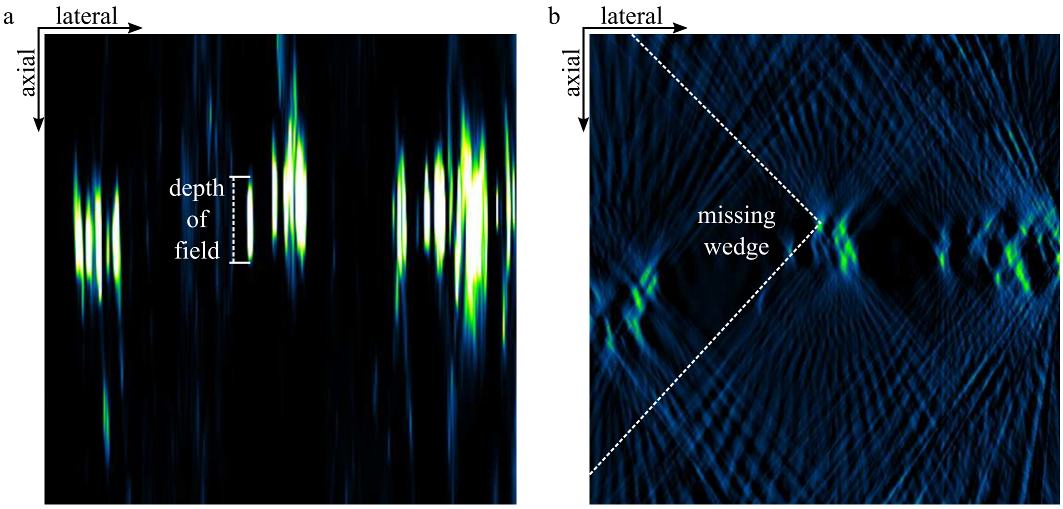

Figure 1.3: a) The dominating artifacts of focal series tomography is an elongation in the axial direction. This elongation corresponds to the depth of field of the microscope. b) In the case of a tilt series with parallel illumination, the missing wedge is clearly visible. A star-like structure of elongated streaks indicate the individual projection directions. Axial elongation is also present but less dominant than in the case of a focal series.

signal-to-noise ratio (SNR) that is a result of electron dose restriction.

In this thesis, the combination of tilt- and focal series is proposed as a method to improve on the issue of axial elongation. The objectives addressed in this thesis include:

- How can the tomographic reconstruction problem be solved for data recorded with a combined tilt- and focal series (CTFS)?
- To what extent can the incorporation of a focal series provide additional resolution in axial direction over the information contained in a tilt series alone? More specifically: Is there a statement comparable to the Fourier slice theorem that applies to a combined tilt- and focal series?
- What are the implications for the theory of tomographic reconstruction when considering images that are recorded with limited depth of field?
- What are practical aspects of combined tilt- and focal tomography? How should a software architecture be designed to handle this kind of data? And finally: What are the limits for the practical applicability?

1.2 Related Work

In the following chapter, previous approaches to 3D STEM microscopy are presented. Hereby, we first present work on the different variations of the acquisition schemes tilt series and focal series. The more exotic approaches "Big Bang" tomography and Ptychography are briefly presented for completeness. After discussing methods to acquire data, we give an overview of methods to process this data, i.e. to tomographic reconstruction. Hereby, we restrict ourselves to approaches directly relevant to this thesis, which are iterative methods for tomographic reconstruction of data acquired with some variation of a tilt series.

1.2.1 Tilt Series Tomography

The primary method currently used for studying the 3D organization of the cellular ultrastructure is tilt-series transmission electron microscope (TEM) (Hoenger & McIntosh, 2009; Kourkoutis, Plitzko, & Baumeister, 2012). A 3D volume is reconstructed from images recorded at several projections obtained by mechanically tilting the sample stage. The resolution is in the range of 2-20 nm, thereby filling a critical length scale between the atomic resolution of X-Ray crystallography (Smyth & Martin, 2000) and single particle electron tomography (Frank, 2006), high-resolution confocal light microscopy (Hell, 2007) (200 nm), and X-Ray microscopy (Meyer-Ilse et al., 2001) (50 nm).

The most commonly used recording scheme is single-tilt tomography. In this scheme, the specimen is rotated around a single axis that is perpendicular to the beam direction. The tilting takes place in constant increments over an angular range. However, the tilt range strongly influences the resolution of the 3D reconstructions (Koster et al., 1997; Fernandez, 2012) and is usually restricted to about ±70° as explained before. The tomographic reconstruction then suffers from missing information and limited resolution on account of this so-called "missing wedge" (Figure 1.4a). A quantitative relationship between the absence of certain frequency components and the missing wedge is detailed later in Section 2.2, Equation 2.6.

Tilt Series Alignment

One issue with tilt series based tomography is the misalignment of tilt series because of mechanical imprecisions in the stage rotation. Here, misalignment means that besides the intended rotational movement around the tilt axis, the sample stage undergoes additional, random movements such as lateral translations, magnification changes, and rotations around different axis. Alignment errors are typically addressed by first computationally estimating the error and then re-projecting the individual images of a tilt series to form a fully aligned stack. Alternatively, the projection parameters related to each image could be stored and considered in the reconstruction algorithm. Either way, the computationally hard task is to estimate the alignment errors.

The task of alignment can be simplified by introducing a dispersion of markers such as gold nanoparticles into the specimen (Pennycook & Nellist, 2011, p. 364). The movement of these particles in each projection can be tracked, and misalignments can be determined by least-squares tracking of fiducial with comparison to a reference projection (Berriman, Bryan, Freeman, & Leonard, 1984; Lawrence, 1984; Olins et al., 1983). The basic method has been improved incrementally, for example by more selective marker detection algorithms based on second-order derivatives (Cao, Takaoka, Zhang, & Nishi, 2011), Markov random fields (Amat et al., 2008) and most recently using optical flow methods (Abrishami et al., 2015).

In situations where the use of markers is prohibited for reasons in the sample preparation or because the resulting artifacts are undesirable, markerless methods can be used. The most common approach is to detect features, "landmarks" in the images and use those positions as markers (Winkler & Taylor, 2006; Sorzano et al., 2009).

The computational task of tilt series alignment is remarkably closely related to the problem known as "bundle adjustment" in the computer vision community that refers to the alignment of camera systems from several photographs. As drastic advancements have recently been reported in this community (e.g. Jian, Balcan, & Dellaert, 2012), it might be worthwhile to investigate if methods recently introduces in the computer vision community can be transfered to electron tomography.

Alternative Tilt Geometries

As mentioned before, one of the two main limiting factors of electron tomography is axial elongation. A number of different solutions have been explored as explained in the following. The missing wedge can be reduced to a missing pyramid using double-tilt tomography (Penczek, Marko, Buttle, & Frank, 1995; Mastronarde, 1997). In this recording scheme, two separate tilt series are recorded with both tilt series in perpendicular direction. Thus, if the beam direction is the z direction, a double-tilt series can be realized by first rotating

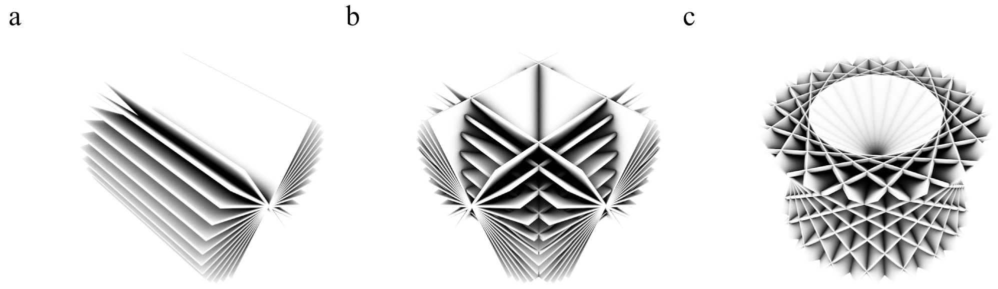

Figure 1.4: Comparison of the different tilt geometries. a) Single tilt results in a “missing wedge”. b) In the double-tilt geometry, the missing wedge is reduced to a “missing pyramid”. c) In conical tilt geometry, the missing region has the shape of a double cone. Figure adapted from (Lanzavecchia et al., 2005).

from ±70° around the x-axis and then by ±70° around the y-axis to generate a single dataset (Figure 1.4b).

Another recording scheme is conical tomography (Lanzavecchia et al., 2005). Hereby, the specimen is first rotated by a fixed angle around an axis in the lateral plane. The tilt series is then generated by successively tilting around the beam axis. Figure 1.4c depicts the principle. In this recording scheme, the missing wedge is reduced to a missing cone, leading to an isotropic resolution in the xy plane. The scheme is therefore particularly beneficial for the study of thin structures such as channels, receptors, and transporters in biological membranes.

Impact of Sample Thickness

The impact of missing vertical information is more severe for imaging samples thicker than several mean free path lengths for electron scattering, typically a few hundreds of nanometers for biological samples. Beyond that, the resolution of TEM tomography is reduced by electron-matter interactions. Firstly, (multiple) elastic scattering events lead to an angular broadening of the electron beam, especially in areas behind high-density objects. Secondly, inelastic scattering spreads the energy spectrum of the electron beam leading to chromatic blurring of the TEM objective lens (Reimer, 1998). Chromatic blurring in thick biological samples can be reduced by introducing energy filtering (Koster et al., 1997) or chromatic aberration correction (Baudoin, Jerome, Kübel, & de Jonge, 2013).

1.2.2 Focal Series Tomography

Avoiding tilting altogether is also possible. By recording of focal series to obtain 3D STEM information (Behan, Cosgriff, Kirkland, & Nellist, 2009; Borisevich, Lupini, & Pennycook, 2006; de Jonge, Sougrat, Northan, & Pennycook, 2010; Dukes, Ramachandra, Baudoin, Gray Jerome, & de Jonge, 2011; Frigo, Levine, & Zaluzec, 2002). Hereby, axial information is retrieved from a stack of images with different focal values. The feasibility of focal series STEM tomography is tightly coupled to the development of spherical aberration correction (Krivanek, Dellby, & Lupini, 1999), as the corrections allow a strongly limited depth of field as a side effect. The 3D data set is deconvolved to remove the blurring effect of the limited DOF (Ramachandra & de Jonge, 2012). But the focal series data suffers from a vertically elongated point spread function, and the vertical resolution is even further reduced by shadowing effects below strongly scattering objects (Behan et al., 2009; de Jonge et al., 2010).

1.2.3 Exotic Sources of Three Dimensional Resolution

"Big Bang" Tomography

The method called "Big Bang" tomography (Van Dyck, Jinschek, & Chen, 2012) is a 2.5D reconstruction scheme that works on individual projections. The method is capable of reconstructing the positions of individual atoms of very thin foils such as single-layer and double-layer graphene with precision of several picometer. The first step of the method is reconstructing the exit wave, i.e. generating amplitude and phase images in separate channels of the projection. This can be achieved using focal series reconstruction (Coene, Thust, Op de Beeck, & Van Dyck, 1996; Hsieh, Chen, Kai, & Kirkland, 2004), off-axis holography (Lehmann & Lichte, 2002), or phase plates (Van Dyck, 2010).

Atom positions are now reconstructed one by one where the xy-positions are treated separately from the z-positions. The methods assumes that the specimen foil is thin enough that the individual atoms are clearly separated in a projection perpendicular to the foil extent. Atom positions in xy-direction are reconstructed by 2D fitting of an ideal lattice.

In order to reconstruct z-positions, the local area around the projection of each atom is separated and transfered to Fourier space. The phase difference to the unscattered wave is determined per Fourier component, and plotted over spatial frequency. This diagram named "hubble plot" originates in astrophysics and is responsible for the name of the tomographic method. By exploiting a linear relationship between phase difference and the distance to the origin of the unscattered wave, the z-positions of the individual atoms can be reconstructed.

Ptychography

The basic idea of ptychography (Hüe, Rodenburg, Maiden, Sweeney, & Midgley, 2010; Hüe, Rodenburg, Maiden, & Midgley, 2011) is to remove the need for an objective lens from the image formation process and replace the optics by diffraction imaging and software (Humphry, Kraus, Hurst, Maiden, & Rodenburg, 2012). The experimental setup uses a STEM microscope together with a charge-coupled device (CCD) detector. The electron beam is moved out of focus such that a small patch of the specimen is illuminated. The CCD detector is placed in the near field of the beam and the resulting refraction pattern is recorded. By scanning the electron beam in relatively large steps over the specimen, a sequence of diffraction patterns is recorded, each representing a different patch of the specimen. The first step of the reconstruction algorithm is to solve for the phase of the diffraction pattern scattered by the object (Nellist, McCallum, & Rodenburg, 1995). The reconstructed wave is then propagated to the specimen, which at this stage is modeled as a plane (Humphry et al., 2012). The plane model of the specimen can be replaced by a multi-slice model such that axial resolution can be generated (Maiden, Humphry, & Rodenburg, 2012). However, the study was conducted using an optical microscope and to the author's knowledge, 3D STEM ptychography has not yet been demonstrated successfully.

1.2.4 Reconstruction Methods

A large number of iterative methods exists to solve the reconstruction problem for data acquired by tilt series electron microscopy. The first iterative method proposed was the algebraic reconstruction technique (ART) (Gordon, Bender, & Herman, 1970). ART is a sequential implementation of the Kaczmarz algorithm (Kaczmarz, 1937) assuming a line model for the forward projection and pixel basis functions. Subsequently, proposed algorithms include sequential iterative reconstruction technique (SIRT) (Gilbert, 1972a), which is an implementation of the Landweber iteration (Landweber, 1951) and simultaneous algebraic reconstruction technique (SART) (Andersen & Kak, 1984), which is a block iterative version of ART and makes the Kaczmarz algorithm accessible for parallel implementations. All those algorithms are regularization algorithms in the sense that they converge to the minimum of the L2-norm of the residual (Gordon et al., 1970; Norton, 1985; Jiang & Wang, 2003).

The algorithms mentioned above follow a common scheme and differ mainly in their update strategy as follows. In ART, the solution is refined pixel-by-pixel, i.e. the forward projection of one pixel is computed and the volume is corrected with the according back projection before moving to the next pixel, which makes the method a row-action algorithm. In SIRT, the forward projection of all projections are computed before correcting the volume. In SART, forward projections are computed for one projection direction at a time. This scheme was formalized in (Censor, 1990) by the introduction of variable block algebraic reconstruction techniques as a common group of algorithms. Convergence of variable block algorithms can be proven with relatively few assumptions of the system matrix (Jiang & Wang, 2003; Wang, Zheng, & Member, 2007; Yan, 2010), leading to robust proofs that are often applicable even after a slight modification of the algorithm.

Much effort has been spent on efficient implementations of the methods on different hardware platforms. Most relevant for this thesis is work with a focus on implementations on graphics processing unit (GPU). A first study using the CUDA environment was presented by (Scherl, Keck, Kowarschik, & Hornegger, 2007), more thorough performance measurements were done subsequently in (Castano Diez, Mueller, & Frangakis, 2007; Keck, Hofmann, Scherl, Kowarschik, & Hornegger, 2009; W. Xu et al., 2010). An investigation of low-level accelerations techniques and their impact on reconstruction performance was presented in (Palenstijn, Batenburg, & Sijbers, 2011). A different implementation approach using sparse matrix algebra was proposed in (Vazquez, Garzon, & Fernandez, 2011). Opposed to the belief that highest reconstruction performance requires the use of GPU, in (Agulleiro & Fernandez, 2011), an optimized implementation using single instruction multiple data (SIMD) instructions on multi-core central processing units (CPUs) was presented and it was claimed that this approach outperforms GPUs implementations.

While improvements on the implementation can lead to faster reconstruction times, higher image quality, and resolution requires modification of the reconstruction algorithm, typically by improving the assumptions on the forward projection. A typical assumption is that the images acquired from the microscope are subject to Gaussian noise. With the very tight restrictions on electron dose for many biological specimen, this assumption might not be very accurate. Therefore, the maximum likelihood method is based on the assumption that the noise in the images is Poisson noise, not Gaussian. The method was introduced in (Dempster, Laird, & Rubin, 1977) and implemented in the field of electron tomography in (Coene et al., 1996).

A rather trivial assumption often used for X-ray tomography that is usually invalid for electron tomography, is that the object function is zero outside the reconstruction volume, i.e. that the sample fits entirely into the illuminated area of the microscope. For reasons in the sample preparation, most electron tomography samples have the shape of thin foils and thus are better approximated by an infinite slab relative to the illuminated region. This leads to artifacts in the outer regions of the tomogram, unless corrected by the method "long object compensation" (W. Xu et al., 2010).

A large potential for improved methods lies in a more physically accurate modeling of the image formation and radiation propagation. Important examples of this research direction include correction for the contrast transfer function for off-focus phase contrast imaging in TEM (Penczek, Zhu, & Schröder, 1997; Voortman, Stallinga, Schoenmakers, van Vliet, & Rieger, 2011) and a correction for Bragg scattering (Venkatakrishnan, Drummy, De Graef, Simmons, & Bouman, 2013).

An entirely different research direction is changing the regularization, i.e. not searching the minimum of the L2-norm of the residual but a solution that is optimal under a different norm. The idea of total variation minimization (TVM) in compressed sensing (L. I. Rudin, Osher, & Fatemi, 1992) in the context of tomography was first proposed in (Donoho, 2006) and further investigates in (Goris, Van den Broek, Batenburg, Heidari Mezerji, & Bals, 2012). A combination of maximum likelihood methods and TVM was proposed in (Yan & Vese, 2011). The TVM approaches generally aim to optimize the L1-norm instead of the L2-norm, thus changing the regularization of the problem. TVM approaches find solutions that have sparse gradients, i.e. exhibit regions with constant intensity values and sharp edges, which is often the case particularly for samples from the material sciences.

For reasons of completeness, direct inversion methods such as convolution back projection (Bracewell & Riddle, 1967; Ramachandran & Lakshminarayanan, 1971; Gilbert, 1972b) and weighted back projection (e.g. Radermacher, 1992) should be mentioned as well. Those methods are based on the idea of expressing the imaging process as a mathematical transform. The transform is inverted analytically and the inverse transform is computed using numerical methods. Direct inversion methods have enormous importance in the field of medical X-Ray CT because of their high computational efficiency. However, they tend to perform poorly under conditions typically encountered in electron tomography, such as low SNR and incomplete projection set (Gordon et al., 1970). Furthermore, it is not obvious how the STEM transform (Section 2.4.2) can be incorporated in a direct inversion method. Therefore, this algorithmic direction is considered out of scope for the purpose of this thesis.

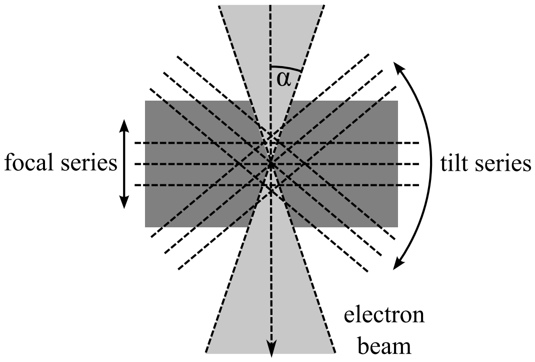

Figure 1.5: Schematic overview of the CTFS recording scheme. The specimen was rotated in relatively large tilt increments over the possible tilt range but for every tilt direction, a through-focal series was recorded. Figure from (Dahmen et al., 2014a).

In summary, several approaches exist to aquire 3D information at the nanoscale. In tilt series based transmission electron microscopy, the specimen is rotated mechanically and images are acquired from different directions, typically using parallel illumination. Iterative reconstruction algorithms are used to reconstruct the tomogram. Alternatively, a focal series can be used to generate 3D information from images with limited depth of field. In this case, a deconvolution is used to remove the blurring of those parts of the specimen that are out of focus. However, both approaches suffer from severe axial elongation artifacts. In the case of the tilt series, they originate from the missing projection set. In the case of focal series tomography, the depth of field is still several orders of magnitude too large to achieve isotropic resolution in all dimensions.

1.3 Own Contributions

The CTFS as a new recording scheme for STEM imaging was introduced (Dahmen et al., 2014a, 2014b, 2014c). In this technique, the specimen is rotated in relatively large tilt increments over the possible tilt range but for every tilt direction, a through-focal series is recorded (Figure 1.5). The main contributions of this work are:

- The CTFS as a new recoding scheme for the image acquisition.
- A new method for axial alignment of confocal STEM images.
- The tilt- focal algebraic reconstruction technique (TF-ART) as a new algorithm for tomographic reconstruction of STEM images.

One drawback of the method was that experimental results showed slow convergence that required ≈ 120 iterations to reach optimal reconstruction results. This stands in contrast to known results from SART reconstructions of pure tilt-series dataset, that typically converge after only 3-5 iterations. This huge difference raises both theoretic and practical concerns about the choice of the system matrix and possible improvements. Therefore, a different back projection operator with better theoretic justification was therefore derived analytically and implemented (Dahmen, Kohr, de Jonge, & Slusallek, 2015). It was shown experimentally that this new back projection drastically improves the convergence characteristics of the method. Main contributions include:

- The formalization of the STEM-transform as a linear operator on functions defined on $\mathbb{R}^3$.
- Proof that the STEM transform is a linear convolution, self-adjoint and a generalization of the ray transform.
- A new reconstruction method based on the adjoint STEM transform with drastically improved convergence characteristics.

Besides the large number of existing software packages for electron tomography, it was perceived that all existing implementations fail to solve some of the fundamental requirements from an architectural point of view. Therefore, a new software package for tomographic reconstruction was therefore presented (Dahmen, Marsalek, et al., 2015). The software is called “Ettention”. The work addresses the issues by creating a general set of high-performance GPU primitives, building blocks, for quickly assembling situation-specific advanced iterative reconstruction algorithms.

1.4 Thesis Structure

This document is structured as follows: after an introduction in Chapter 1, the thesis starts with a theoretic part in Chapter 2. Here, the STEM transform is introduced as a linear operator. It is shown that the STEM transform is a linear convolution and a generalization of the ray transform for parallel illumination that considers imaging with a limited DOF. The Fourier transform of this operator is considered, which can be interpreted geometrically in a somewhat surprising way. Additionally, the adjoint operator of the STEM transform is derived by showing that the operator is self-adjoint.

In Chapter 3, we present an iterative algorithm for solving the tomographic reconstruction problem for combined tilt- and focal series. The algorithm uses an implementation of the STEM transform based on stochastic ray tracing as a forward model. As a back projection, two different methods are implemented and compared. The first method, called “unmatched” back projection is based on a heuristic weighting factor while the second method, called “matched” back projection uses an efficient implementation of the adjoint operator derived in Chapter 2. The method is evaluated on experimental data and we show that the axial elongation artifacts can indeed be improved significantly. We compare convergence performance of both back projections experimentally and it is shown that the matched back projection converges drastically faster than the unmatched back projection.

In Chapter 4, we propose a software architecture for the problem “image reconstruction from projections” in a more general context. A toolbox of building blocks for iterative reconstruction algorithms is presented that separates most of the technical aspects of efficient programming on parallel architectures from the problem domain. It is shown how the toolbox can be used to implement the method presented in Chapter 3, but also how it can be generalized and applied to a wide range of different tomographic reconstruction problems.

Chapter 2

Theory

2.1 Definition of the Term "Resolution"

The resolution of the human eye is defined as the smallest possible distance at which two points are still distinguishable. The unaided human eye has a resolution of about 0.2 *mm* (Russ, 2006). In the context of microscopes, the resolution (or resolving power) of a microscope is defined as the closest distance two points can have and still be distinguishable using the microscope, assuming ideal conditions.

2.1.1 Resolution Limit by the Rayleigh Criterion

When a wave passes through an opening in a barrier, such as an aperture in a lens, it is diffracted by the edges of the aperture. Even a perfectly shaped lens will be limited in its resolving power by this diffraction. A high quality optical lens is referred to as a diffraction-limited lens. Any further effort to improve the quality of the lens surface will not improve its resolution (Born & Wolf, 1997).

The Rayleigh criterion (Figure 2.1) for visible light microscopes states that the smallest distance that can be resolved, σ, is given by

$$\sigma = \frac{0.61\lambda}{\mu \sin\gamma} = \frac{0.61\lambda}{NA} \approx 0.61\lambda \tag{2.1}$$

In this equation, μ is the refractive index of the viewing medium and γ is the semi-angle of collection of the magnifying lens. $\mu \sin\gamma = NA$ is called the

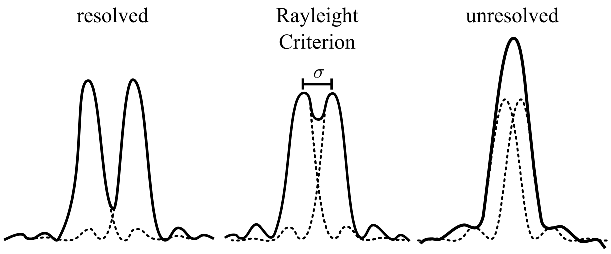

Figure 2.1: The Rayleigh criterion: Two points are regarded as just resolved when the principal diffraction maximum of one image coincides with the first minimum of the other.

numeric aperture and is a property of the microscope (≈ 0.1-1.4 for optical microscopes). The Rayleigh criterion translates from visible light microscopy to electron microscopes, where the numeric aperture is approximately one (Williams & Carter, 2009). As a result, the upper limit for the achievable resolution is roughly half the wavelength.

Visible light has a wavelength of 400-700 nm, effectively limiting the resolution of visible light microscopes to approximately 250 nm (Abbe, 2004). Approaches to achieve resolutions better than this using optical light exists, and are referred to as super resolution microscopy (Leung & Chou, 2011).

A different approach is using a particle with magnitudes smaller wavelength. The wavelength of an electron depends on its energy and is given by

$$\lambda = \frac{h}{p} = \frac{h}{m_0 e U} \tag{2.2}$$

where m_0 is the resting mass of the electron, e is the electric charge, and U is the voltage used for the acceleration. It is noticeable here that the speed of an electron is a function of the total potential difference alone and does not depend on the geometric setup of the accelerating aparatus within wide tolerances.

With typical acceleration voltages between 80 kV and 300 kV, electron wavelength lies between 2.4 pm and 4.4 pm giving a maximal spacial resolution of 1.2 pm to 2.2 pm. This limit is not reached by far (Krivanek et al., 1999; Haider, Uhlemann, & Zach, 2000; Bleloch & Lupini, 2004) for technical reasons such as chromatic aberrations and spherical aberrations which are beyond the scope of this thesis.

2.1.2 Resolution Limit by Nyquist-Shannon Theorem

The Nyquist-Shannon theorem states, that if a function $p(x)$ of space, the signal, contains no spacial frequencies higher than B, i.e. the Rayleigh criterion holds. This means $p(x)$ it is completely determined by giving its values at a series of points spaced $d_{sampling}$ apart, where

$$d_{sampling} = \frac{1}{2B} \tag{2.3}$$

A microscope is considered with a sensor pixel size of d_{pixel} and an overall magnification of M. By inserting in Equation 2.3 one obtains an upper limit for the spacial frequency which is completely determined.

$$B_{max} = \frac{M}{2\,d_{pixel}} \tag{2.4}$$

For STEM microscopes, the pixel size is determined by the accuracy of the deflector coils that scan the electron beam over the sample, such that the sampling distance can be chosen almost arbitrarily.

However, the lateral resolution is limited not only by the sampling distance of the microscope, but also by the highest frequency the microscope can transmit. The signal can generally be described as the specimen function, convoluted by a function called the point spread function (PSF) that describes the imaging characteristics of the microscope and basically corresponds to the shape of the electron beam. Imaging an hypothetical, infinitesimal small object would generate an image of the PSF, hence the name. The convolution of two band limited functions contains no frequencies that are not present in both functions. Mathematically, this can easily be understood by remembering that the convolution corresponds to a multiplication in Fourier space. As a consequence, choosing a sampling distance smaller than the size of the PSF does not result in an additional gain of information, only in excessive sampling of a band limited functions.

2.1.3 The Full-Width-at-Half-Maximum Criterion

One common method to evaluate resolution is to experimentally determine the PSF. To do so, an image of an approximate point like object, for example a very small nanoparticle, is recorded. Hereby it is crucial to select an object that is known in advance to be smaller than the PSF. The intensity is recorded along a line through the center of that object. The full-width-at-half-maximum (FWHM) resolution is defined as the width of the peak at the value between the background intensity level and the signal maximum (Figure 2.2).

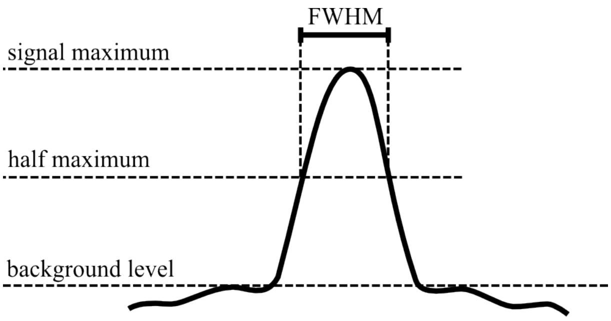

Figure 2.2: Definition of the full-width-at-half-maximum resolution. An approximate point-like object is imaged and an intensity plot is created through the center of the image. The signal background level is determined as well as the maximum peak. The peak diameter is then measured at the half-maximum between the two values.

2.1.4 Contrast and Electron Statistics

In the above considerations, it is stated that a sensor samples a signal, without further considering the physical nature of that signal. The exact physical properties (such as mass density or atomic number) of the specimen that relate to the signal depend on the imaging mode. However, in the context of STEM, it is always correct to state that the signal corresponds to the probability of an incident electron reaching (positive signal) or not reaching (negative signal) the detector pixel. So the signal is a function $g(u)$ that relates a spatial position to a probability. When it is said that the signal $g(u)$ is sampled, the detector records samples of this probability variable. Hereby, each electron works as a binary sample which takes values one in case the electron reached the detector or value zero in case it did not reached the detector (Figure 2.3). In the case of the conventional transmission electron microscope (CTEM), where an entire image is recorded simultaneously, the situation is slightly more complicated because an electron that is scattered and thus misses an detector element might hit the element of a different pixel.

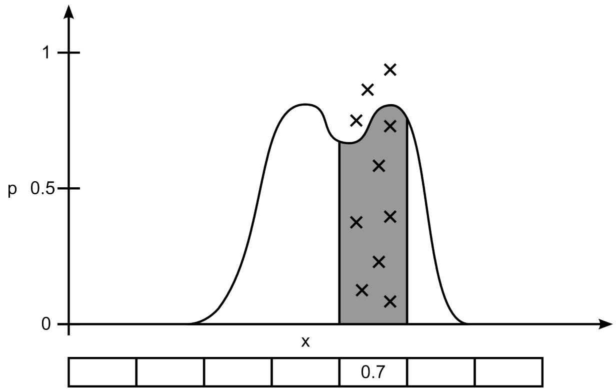

Figure 2.3: The signal is the probability of any electron reaching the detector, expressed as function of location. When the signal is sampled, the detector measures the pixel value counting how many of a given number of electrons reached the detector during the dwell time of a pixel.

Still, the sampling obviously gives the correct signal value in the limit. But assuming the pixel is exposed to a finite number of electrons N, the estimate is correct only within a certain tolerance which is given by

$$SE = \frac{1}{2\sqrt{N}}. \tag{2.5}$$

This formula implies that the precision of the estimate can be increased arbitrarily by increasing N, i.e. spending more electron dose per pixel. On the other hand, if a fixed electron dose is assumed, increasing the spacial resolution by using smaller detector elements also decreases N. Assuming the presence of frequencies beyond the Nyquist frequency, dividing the pixels distance d_{pixel} by a factor of two will improve the resolution by a factor of two as well. Because the detector element receives electrons proportional to its area (not edge length), the number of primary electrons per pixel is divided by a factor of four and the standard error of the pixel gray value is doubled.

To summarize, STEM images are subject to noise, which in this context is a fundamental consequence of imaging with a limited number of electrons per pixel. Improved noise can be traded for decreased resolution by down sampling the image using standard image processing methods. Improving the product of noise and resolution, however, is limited by the acceptable electron dose, which depends on factors such as the type and preparation procedure of the specimen, the intended application, and the acceleration voltage and is a fundamental limiting factor to electron tomography. For this thesis, this implies that all following considerations are made under the assumption that the image signal suffers from a low SNR and that this noise translates to imaging errors in the tomograms.

2.1.5 Resolution in Three Dimensions: Isotropy and Axial Elongation

So far, resolution has been considered in 2D only. Hereby, the orientation of the two dimensions is arbitrarily interchangeable because the imaging process can freely be rotated around the axis given by the beam direction. This implies that there is no principal anisotropic component in 2D images from electron microscopes. When considering resolution in 3D, this situation changes because the third direction is always resolved by means of tomographic reconstruction, as opposed to the scanning of the electron beam. As a consequence, the tomogram has two interchangeable lateral dimension and one fundamentally different, axial direction. So 3D volumes do have an anisotropic component for principal reasons and the axial direction typically has lower resolution than the lateral directions.

In order to quantify the loss of resolution in axial direction, a measure based on the FWHM is typically used. A line of arbitrary direction in the lateral plane is considered through the center of a small nanoparticle and the FWHM in lateral direction is measured as explained in Section 2.1.3. A second line in axial direction through the center of the same nanoparticle is used to measure the FWHM in axial direction. The axial elongation factor e_{xz} is defined as the ratio of the two resolution values. For perfect isotropic resolution, e_{xz} equals one.

The axial elongation factor can be generalized to the concept of an angle dependent elongation factor e_γ if the second resolution value is measured not in axial direction but in an arbitrary direction that forms an angle γ with the beam direction. The angle dependent elongation factor is one for $\gamma = \pi/2$ by definition and equals e_{xz} for $\gamma = 0$. By plotting e_γ over γ, the loss of resolution in different directions can be characterized.

2.2 The Ray Transform for Parallel Illumination

In the context of tilt series electron tomography, the computational problem of generating 3D information, i.e. information in the axial direction is called "tomographic reconstruction". In the following, an overview of the continuous formulation of the tomographic reconstruction problem is given, roughly following the presentation in (Natterer & Wübbeling, 2001). Here, the volume is modeled as a function $f(x) : \mathbb{R}^3 \to \mathbb{R}$ that maps points in space to a density value and similarly, the images are represented as a function $g(x) : \mathbb{R}^2 \to \mathbb{R}$ that map points in the signal space to gray values. The imaging process is modeled as a transform $\mathcal{P}$ that integrates h over straight lines.

Each line is represented by a direction $\theta \in S^2$ and a point $u \in \theta^\perp$ on the plane perpendicular to θ as $\{u + t\theta : t \in \mathbb{R}\}$. Then the ray transform for parallel illumination (sometimes called X-Ray transform for historic reasons) $\mathcal{P}_\theta$ can be defined as

$$\mathcal{P}_\theta f(u) = \int_{\mathbb{R}} f(u + t\theta)\,\mathrm{d}t, \quad u \in \theta^\perp. \tag{2.6}$$

In the slightly different 2D case, i.e. when a 2D image is reconstructed from one-dimensional (1D) projections, the ray transform equals the famous Radon transform except for notation. However, the case of particular interest for this thesis is 3D tomography, i.e. the reconstruction of a 3D volume from 2D projections, so the representation in Equation 2.6 will be used.

The most important result that is directly based on this representation is the 3D version of the Fourier slice theorem, which is given by

$$\mathcal{F}((\mathcal{P}_\theta f)(\xi)) = (2\pi)^{1/2}\mathcal{F}(f(\xi)). \tag{2.7}$$

for $\xi \perp \theta$. The theorem states that the Fourier transform of a projection contains the frequencies in the plane g_θ through the origin and orthogonal to the projection direction θ (Figure 2.4). A proof of the theorem along with additional properties of the ray transform for parallel illumination can be found in (Natterer & Wübbeling, 2001), Chapter 2.2.

2.3 Tomographic Reconstruction Methods

In the continuous formulation, reconstruction algorithms, i.e. algorithms to solve the tomographic reconstruction problem, approximate a solution to the system

$$[\mathcal{A}_k f](\boldsymbol{u}) = g_k(\boldsymbol{u}), \quad k = 1, \ldots, K. \tag{2.8}$$

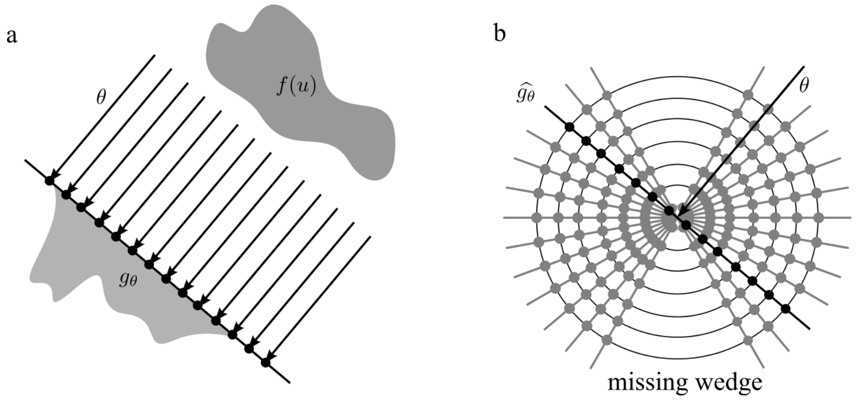

Figure 2.4: Schematic representation of the Fourier slice theorem. a) An object function $f(u)$ is projected in direction θ resulting in the image function g_θ. b) The Fourier transform of g_θ corresponds to a single slice in Fourier space, orthogonal to θ.

The tomogram is modeled by an unknown function $f : \mathbb{R}^3 \to \mathbb{R}$ which relates a position $\boldsymbol{u} \in \mathbb{R}^3$ in the sample to a density value $f(\boldsymbol{u})$. Likewise, the images from the microscope are represented by functions $g_k : \mathbb{R}^2 \to \mathbb{R}$ relating a pixel coordinate in the images to a gray value or an intensity. Note, however, that in the CTFS acquisition scheme, each g_k consists of a stack of images, i.e. comprises a 3D function. Nevertheless, the structure of the reconstruction scheme remains the same. The imaging process is modeled by a collection of linear operators $\mathcal{A}_k$ establishing a mathematical connection between a tomogram f and the corresponding projection data g_k, i.e. a model for the physics of image formation.

For processing on a computer, a discrete representation of f and g_k is required. This discretization is reached by decomposing these functions according to

$$f(\boldsymbol{u}) \approx \sum_{i=1}^{N} a_i(\boldsymbol{u}) X_i \quad \text{and} \quad g_k(\boldsymbol{u}) \approx \sum_{j=1}^{M} b_j(\boldsymbol{u}) B_{k,j} \tag{2.9}$$

with respect to basis functions a_i and b_j (Figure 2.5). Typical choices are pixel functions, piecewise linear functions or “blobs” (Marabini, Herman, & Carazo, 1998). Hereby, the choice of basis has a different motivation for the voxel of the reconstruction volume and the pixels in the images. For the volume, it is difficult to define how "ideal“ basis functions should look like, and the basis can be chosen under mathematical considerations to find a good discrete approximation of the objective function f. Typical choices are pixel functions for their ease of implementation or blobs for their band-limiting properties. The choice of basis functions for the image pixels however is tightly coupled to the physical nature of the detector. This is the case because the image is acquired from discrete measurements in the first place so the ideal basis functions is the one that perfectly reflects the properties of the detector element. In the case of STEM imaging, the ideal basis is the δ-function, as typically one pixel corresponds to exactly one beam position. In the case that subpixel scanning is used, i.e. the beam is scanned over the area of the pixel during the pixel dwell time, the basis function should be replaced by a pixel function.

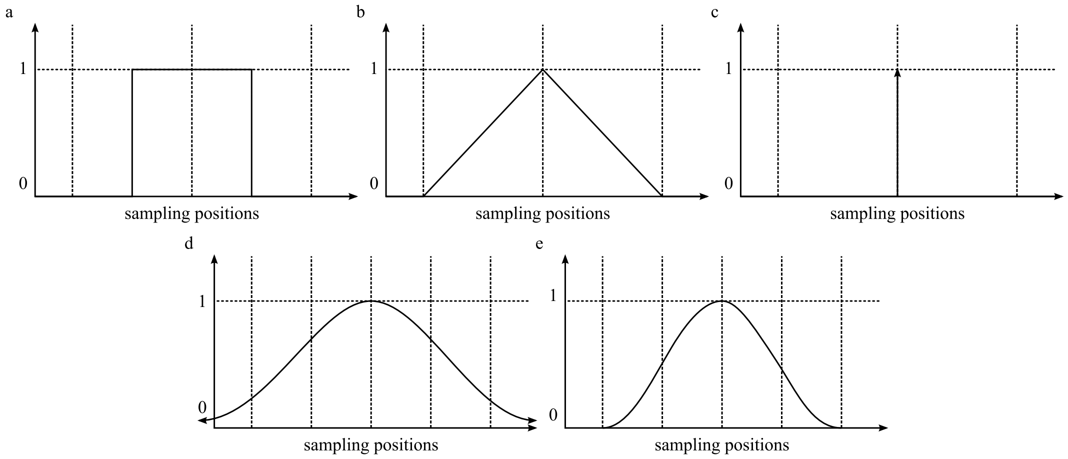

Figure 2.5: An overview of possible choices for the basis functions. a) Pixel (piecewise constant) basis functions. b) Piecewise linear basis functions. c) δ-function. d) Gaussian. e) Blobs.

The decomposition mentioned above turns Equation 2.8 into the linear system

$$A_k X = B_k, \tag{2.10}$$

where the discrete representation of the observed data in the k'th image is B_k and X stands for the discretization of the tomogram. The matrix A_k is therefore the discrete representation of the linear operator $\mathcal{A}_k$ in the chosen bases a_i and b_j.

In the same sense, the adjoint operator $\mathcal{A}_k^*$, often called back projection and defined by

$$\langle \mathcal{A}_k f, g_k \rangle = \langle f, \mathcal{A}_k^* g_k \rangle, \tag{2.11}$$

is represented by the transposed matrix A_k^T, i.e. the transposed matrix is the backprojector consistent with the forward operator $\mathcal{A}_k$ and the choice of bases. The exact meaning of Equation 2.11 in terms of square-integrable functions will be clarified in Section 2.4.5.

Kaczmarz algorithm is an iterative scheme consisting of an inner iteration cycling through all images in the dataset and an outer iteration which simply stands for a repetition of the inner loop. Starting with an initial guess $f = f^{(0)}$ (often $f^{(0)} = 0$), the k'th step in the inner loop, corresponding to image k, reads as

$$f \leftarrow f + \Lambda \mathcal{A}_k^* (\mathcal{A}_k \mathcal{A}_k^*)^{-1} \big(g_k - \mathcal{A}_k f \big) \tag{2.12}$$

or, in the discrete formulation

$$X \leftarrow X + \Lambda A_k^T (A_k A_k^T)^{-1} \big(B_k - A_k X \big). \tag{2.13}$$

These formulas can be interpreted as follows: first, the current guess (f or X) is projected forward and subtracted from the actual data (g_k or B_k), resulting in the so-called *residual*. In an intermediate step, the inverse of $\mathcal{A}_k \mathcal{A}_k^*$ or $A_k A_k^T$, respectively, is applied to the residual. Finally, the intermediate result is projected back via $\mathcal{A}_k^*$ or A_k^T, yielding the update to be added to the current guess. The factor $\Lambda \in (0, 2)$ acts as a relaxation parameter. Once all data in the dataset have been processed, the inner iteration starts over, and the outer iteration number is raised by one. The iteration continues until some termination criterion is met, typically for a fixed number of iterations or until the residual error drops below some user defined threshold.

A projection-backprojection pair that implicitly defines the matrices A_k and the corresponding transposes A_k^T is an essential component of the algorithm. The projection should model the imaging geometry and the physics of the imaging process as accurately as possible, while according to the Kaczmarz algorithm, the back projection should be the adjoint of this operator since it represents the last step of the update computation in Equation 2.12. However, as shown by (Zeng & Gullberg, 2000) for the matrix formulation, differing back projections are also possible. If a back projection is not the adjoint of the projection, the pair is called “unmatched”. Convergence in a least-squares sense (yet to a different solution than for the matched pair) can be proven also in the case of an unmatched projection-backprojection pair assuming certain conditions on the back projection (Zeng & Gullberg, 2000).

2.4 The STEM Forward Projection

In the design of tomographic reconstruction algorithms for the field of electron tomography, a typical assumption is that the intensity of a pixel in a projection represents the integral along the line to that pixel. This corresponds to the classical ray transform for parallel illumination as presented in Equation 2.6. The line integral of the volume can be efficiently computed using ray-casting (Levoy, 1990). For each pixel in the forward projection, the ray through that pixel is generated and the volume is sampled at uniform intervals along that ray. Alternatively, conservative line drawing algorithms like a 3D version of Bresenham algorithm (Bresenham, 1965) can be used to compute the exact integral value along the ray. The line model of the electron beam is accurate for the approximately parallel illumination in TEM tomography, or for STEM tomography with a small beam convergence angle leading to a large DOF with respect to the sample thickness and assuming δ-functions for the image basis. For other image basis functions, the pixel footprint needs to be sampled using several rays.

2.4.1 Double Cone Model of the Electron Beam

However, in aberration corrected STEM the electron beam is convergent with a focal depth of typically a few nanometers (Lupini & de Jonge, 2011) as a side effect of spherical aberration correction and the line model of the forward projection is no longer a good approximation. For this reason, once should also take convergence of the electron beam into account (Dahmen et al., 2014a). The probe shape in the forward projection used a model consisting of a double cone, as am idealized model for the PSF. Any lateral cut through this double cone results in a circular disc. In this model, the value of a pixel in the projection equaled the volume integral of the gray values of the voxels inside the double cone, weighted by the local current density. We assume that the probe current is homogeneous within each disc and zero outside. As a consequence, the local current density changes with the reciprocal of the disc area, i.e. is larger close to the focus plane. The beam model is called the *STEM transform* and can be parametrized as shown in Figure 2.6.

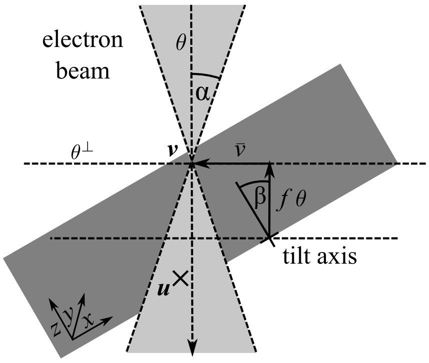

Figure 2.6: Parametrization of the STEM transform. Positions inside the volume are identified by coordinates $u := (x, y, z)$. α is the beam opening semi-angle, β the tilt angle. The unit vector θ denotes the beam direction, the scalar f the focal length, defined as the distance in direction of θ from the tilt axis to the focal plane $\theta^\perp$. $\bar{v} \in \theta^\perp$ is a vector perpendicular to θ, such that the vertex v of the double cone can be expressed as $v = f\theta + \bar{v}$. Figure adapted from (Dahmen, Kohr, et al., 2015).

2.4.2 The STEM Transform as a Linear Operator

In the following, the STEM forward projection is expressed as a linear operator on functions defined on $\mathbb{R}^3$. This representation is useful to derive the corresponding back projection and to investigate analytical properties of the model. In order to simplify notation, all expressions are written without coordinates and are given from the perspective of a fixed specimen and a rotating probe. The experimentally natural perspective of rotating sample and fixed probe can be obtained by applying the geometry transformation $\boldsymbol{u} \mapsto R_{\boldsymbol{\theta}}\boldsymbol{u}$ to integrals, where $R_{\boldsymbol{\theta}}$ is the rotation matrix given in Equation 2.16.

A double cone with an opening semi-angle of $\alpha \in (0, \pi/2)$, its vertex at the origin, and the z-axis as rotational symmetry axis can be parametrized as the set

$$C_\alpha = \left\{\boldsymbol{u} = (\bar{\boldsymbol{u}}, z) \in \mathbb{R}^3 \,\middle|\, |\bar{\boldsymbol{u}}| < |z| \tan\alpha\right\}. \tag{2.14}$$

In order to ensure constant electron flux inside the double cone, a weighting factor w is introduced such that the non-rotated *probe function* or PSF can be written as:

$$p(\boldsymbol{u}) = w(z) \cdot \begin{cases} 1 & \text{if } \boldsymbol{u} \in C_\alpha \\ 0 & \text{else.} \end{cases} \tag{2.15}$$

Here, the weight $w(z) := (\pi\,|z|^2 \tan^2\alpha)^{-1}$ accounts for constant electron flux in that it divides by the area of the lateral cut through the double cone at distance $|z|$ from the focal point, this cut being a disc of radius $|z|\tan\alpha$. The underlying physical approximation is that no electrons are absorbed in the specimen.

To model a projection along the direction vector $\boldsymbol{\theta}$, the symmetry axis is changed to $\boldsymbol{\theta}$ by multiplication with a rotation matrix $R_{\boldsymbol{\theta}}$ that rotates $\boldsymbol{e}_z$ to $\boldsymbol{\theta}$. For instance, in single-axis tilting by an angle β around the y-axis, $\boldsymbol{\theta}$ and $R_{\boldsymbol{\theta}}$ can be explicitly written as

$$\boldsymbol{\theta} = (\sin\beta, 0, \cos\beta), \quad R_{\boldsymbol{\theta}} = \begin{pmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{pmatrix}. \tag{2.16}$$

Rotating the whole set C_α results in the rotated double cone

$$C'_\alpha = \left\{ \boldsymbol{u}' = R_{\boldsymbol{\theta}}\boldsymbol{u} \,\middle|\, \boldsymbol{u} \in C_\alpha \right\}. \tag{2.17}$$

In analogy to the decomposition of a vector $\boldsymbol{u} = (\bar{\boldsymbol{u}}, z)$ into a component z along $\boldsymbol{e}_z$ and a perpendicular vector $\bar{\boldsymbol{u}}$, a rotated vector $\boldsymbol{u}'$ is decomposed into a component $s \in \mathbb{R}$ along $\boldsymbol{\theta}$ and a perpendicular vector $\boldsymbol{\eta}$ by writing

$$\boldsymbol{u}' = s\boldsymbol{\theta} + \boldsymbol{\eta}, \quad \boldsymbol{\theta} \cdot \boldsymbol{\eta} = 0, \tag{2.18}$$

where "$\cdot$" stands for the usual dot product on $\mathbb{R}^3$. Thus, the rotated cone can be written as

$$C'_\alpha = \left\{ (\boldsymbol{u}' = s\boldsymbol{\theta} + \boldsymbol{\eta}) \in \mathbb{R}^3 \,\middle|\, |\boldsymbol{\eta}| < |s| \tan\alpha \right\}. \tag{2.19}$$

The value $g_{\boldsymbol{\theta}}(\boldsymbol{0})$ of a projection of the density distribution f along $\boldsymbol{\theta}$ with focal point $\boldsymbol{v} = \boldsymbol{0}$ is then equal to the integral value

$$g_{\boldsymbol{\theta}}(\boldsymbol{0}) = [\mathcal{A}_\theta f](\boldsymbol{0}) = \int_{C'_\alpha} w(s)\, f(\boldsymbol{u}')\, \mathrm{d}\boldsymbol{u}', \tag{2.20}$$

Finally, in order to place the focal point at $\boldsymbol{v} \in \mathbb{R}^3$ instead of $\boldsymbol{0}$, the set C'_α is shifted by $\boldsymbol{v}$, or equivalently, the volume function f is shifted by replacing its argument with $\boldsymbol{u}' \mapsto \boldsymbol{v} + \boldsymbol{u}'$. This leads to

$$g_{\boldsymbol{\theta}}(\boldsymbol{v}) = [\mathcal{A}_{\theta} f](\boldsymbol{v}) = \int_{C'_\alpha} w(s)\, f(\boldsymbol{v} + \boldsymbol{u}')\, \mathrm{d}\boldsymbol{u}' \\ = \int_{C'_\alpha} w(s)\, f(\boldsymbol{v} - \boldsymbol{u}')\, \mathrm{d}\boldsymbol{u}', \tag{2.21}$$

where the last equality holds since the weight w and the integration domain C'_α are invariant under the variable sign change $\boldsymbol{u}' \mapsto -\boldsymbol{u}'$. In consequence, the forward projection can be expressed as the convolution integral

$$[\mathcal{A}_{\theta} f](\boldsymbol{v}) = \int_{\mathbb{R}^3} p_{\boldsymbol{\theta}}(\boldsymbol{u}') f(\boldsymbol{v} - \boldsymbol{u}')\, \mathrm{d}\boldsymbol{u}' = [p_{\boldsymbol{\theta}} * f](v), \tag{2.22}$$

with the rotated PSF

$$p_{\boldsymbol{\theta}}(\boldsymbol{u}') = p_{\boldsymbol{\theta}}(s\boldsymbol{\theta} + \boldsymbol{\eta}) = w(s) \cdot \begin{cases} 1 & \text{if } \boldsymbol{u}' \in C'_\alpha \\ 0 & \text{else.} \end{cases} \tag{2.23}$$

Furthermore, in order to analyze the relationship between available projection data $g_{\boldsymbol{\theta}}$ and the searched-for density function f, it is important to specify at which positions $\boldsymbol{v} \in \mathbb{R}^3$ the sample can be scanned, i.e. at which locations the focal spot can be placed for one fixed tilt angle. Usually, the focal length, i.e. the focal spot position along the beam, is kept fixed, and only a lateral movement of the beam relative to the sample is performed. In our geometry, this would result in the focal spot $\boldsymbol{v}$ varying over a 2D plane

$$\boldsymbol{\theta}^\perp = \left\{ \boldsymbol{v} \in \mathbb{R}^3 \,\middle|\, \boldsymbol{v} \cdot \boldsymbol{\theta} = 0 \right\} \tag{2.24}$$

perpendicular to the beam direction $\boldsymbol{\theta}$. However, in the CTFS acquisition scheme as discussed in this thesis, a complete *focal series* is recorded at each tilt angle, such that the projection data for one tilt angle is a stack of 2D images, i.e. a 3D data volume. In other words, the projection $g_{\boldsymbol{\theta}}$ corresponding to such a focal series is a 3D function $g_{\boldsymbol{\theta}} : \mathbb{R}^3 \to \mathbb{R}$ defined by the convolution

$$g_{\boldsymbol{\theta}}(\boldsymbol{v}) = [\mathcal{A}_{\boldsymbol{\theta}} f](\boldsymbol{v}) = [p_{\boldsymbol{\theta}} * f](\boldsymbol{v}), \quad \boldsymbol{v} \in \mathbb{R}^3. \tag{2.25}$$

The linear convolution operator $\mathcal{A}_{\boldsymbol{\theta}}$ called STEM transform thus models the acquisition of a focal series for a fixed beam direction $\boldsymbol{\theta}$.

2.4.3 Relation to the Parallel Ray Transform

In the following, we investigate the relationship between the STEM transform and the well-known ray transform for parallel illumination. The volume integral over the double cone as in Equation 2.21 can be expressed as an integral over the collection of lines constituting the cone. These lines are parametrized as follows. For simplicity, the case $\boldsymbol{\theta} = \boldsymbol{e}_z$ is considered first.

The double cone C_α is intersected with the 2D plane $\{(\bar{\boldsymbol{u}}, d) \,|\, \bar{\boldsymbol{u}} \in \mathbb{R}^2\}$ lying at a distance $d > 0$ from the origin. This intersection results in a circular disc of radius $d \tan\alpha$. Every point in this disc uniquely determines a line through the vertex of the double cone, i.e. the integral over C_α can be parametrized as an integral over the circular disc $\{\bar{\boldsymbol{u}} \in \mathbb{R}^2 \,|\, |\bar{\boldsymbol{u}}| < d \tan\alpha\}$ (see Appendix for details), leading to

$$[\mathcal{A}_{\boldsymbol{e}_z} f](\boldsymbol{v}) = \frac{1}{\pi d^2 \tan^2\alpha} \int_{|\bar{\boldsymbol{u}}| < d \tan\alpha} \frac{d}{\sqrt{|\bar{\boldsymbol{u}}|^2 + d^2}} [\mathcal{P}_{\boldsymbol{\omega}(\bar{\boldsymbol{u}})} f](\boldsymbol{v}) \, \mathrm{d}\bar{\boldsymbol{u}}, \tag{2.26}$$

where $\boldsymbol{\omega}(\bar{\boldsymbol{u}}) = (\bar{\boldsymbol{u}}, d)/\sqrt{|\bar{\boldsymbol{u}}|^2 + d^2}$ is the direction vector associated with the line through the point $(\bar{\boldsymbol{u}}, d)$, and $\mathcal{P}_{\boldsymbol{\omega}}$ is the ray transform (Equation 2.2) with direction $\boldsymbol{\omega}$ instead of $\boldsymbol{\theta}$. For an arbitrary beam direction $\boldsymbol{\theta}$, this formula generalizes to

$$[\mathcal{A}_{\boldsymbol{\theta}} f](\boldsymbol{v}) = \frac{1}{\pi d^2 \tan^2\alpha} \int_{\substack{\boldsymbol{\eta} \in \boldsymbol{\theta}^\perp \\ |\boldsymbol{\eta}| < d \tan\alpha}} \frac{d}{\sqrt{d^2 + |\boldsymbol{\eta}|^2}} [\mathcal{P}_{\boldsymbol{\omega}(\boldsymbol{\eta})} f](\boldsymbol{v}) \, \mathrm{d}\boldsymbol{\eta} \tag{2.27}$$

with $\boldsymbol{\omega}(\boldsymbol{\eta}) = (d\boldsymbol{\theta} + \boldsymbol{\eta})/\sqrt{d^2 + |\boldsymbol{\eta}|^2}$. Thus, since the preceding factor $(\pi d^2 \tan^2\alpha)^{-1}$ is the reciprocal of the area of the integration domain, the STEM transform is a (weighted) mean value integral of the ray transforms $[\mathcal{P}_{\boldsymbol{\omega}} f](\boldsymbol{v})$ over all directions $\boldsymbol{\omega}$ inside the double cone.

Geometrically, it is intuitive to assume that for $\alpha \to 0$, the STEM transform should approach the ray transform since the double cone shape of the beam approximates an “infinitely narrow” ray for vanishing opening angle. Indeed, with the help of an integral mean value argument (see Appendix), one can show that

$$\lim_{\alpha \to 0} \frac{1}{\pi d^2 \tan^2 \alpha} \int_{\substack{\boldsymbol{\eta} \in \boldsymbol{\theta}^\perp \\ |\boldsymbol{\eta}| < d \tan \alpha}} \frac{d}{\sqrt{d^2 + |\boldsymbol{\eta}|^2}} [\mathcal{P}_{\boldsymbol{\omega}(\boldsymbol{\eta})} f](\boldsymbol{v}) \, \mathrm{d}\boldsymbol{\eta} = [\mathcal{P}_{\boldsymbol{\theta}} f](\boldsymbol{v}). \tag{2.28}$$

This demonstrates that the double cone model with the chosen weight $w(s)$ is a generalization of the parallel ray model in the sense that the latter is contained as the special case $\alpha \to 0$.

2.4.4 The Fourier Transform of the STEM Transform

In tomographic applications with parallel illumination, the Fourier slice theorem (Equation 2.7) states that the Fourier transform of a projection, that is the Fourier transform of one image in the dataset, contains exactly those 3D spatial frequencies of the searched-for volumetric function which lie on a rotated 2D plane whose normal vector corresponds to the projection axis. With this formula, it became possible to derive an inversion formula for the ray transform by "filling the Fourier space with planes" and by determining the correct weight to account for non-uniform sampling of the Fourier space. For the STEM transform $\mathcal{A}_{\boldsymbol{\theta}}$, a similar result can be derived. Using again the decomposition of a vector $\boldsymbol{\xi}' \in \mathbb{R}^3$ as

$$\boldsymbol{\xi}' = \sigma \boldsymbol{\theta} + \boldsymbol{\zeta}, \quad \boldsymbol{\zeta} \cdot \boldsymbol{\theta} = 0 \tag{2.29}$$

into a component $\sigma \in \mathbb{R}$ along $\boldsymbol{\theta}$ and a perpendicular vector $\boldsymbol{\zeta}$ (see also Equation 2.18), the 3D Fourier transform of a projection $g_{\boldsymbol{\theta}} = \mathcal{A}_{\boldsymbol{\theta}} f$ can be expressed as

$$\widehat{g_{\boldsymbol{\theta}}}(\boldsymbol{\xi}') = \widehat{f}(\boldsymbol{\xi}') \cdot \frac{4}{|\boldsymbol{\zeta}| \tan \alpha} \begin{cases} \sqrt{1 - \frac{\sigma^2}{|\boldsymbol{\zeta}|^2 \tan^2 \alpha}} & \text{if } |\sigma| < |\boldsymbol{\zeta}| \tan \alpha \\ 0 & \text{else,} \end{cases} \tag{2.30}$$

as shown in (Intaraprasonk, Xin, & Muller, 2008, Equation 27). This formula can be interpreted in a way similar to the classical Fourier slice theorem. The Fourier transform of $\mathcal{A}_{\boldsymbol{\theta}} f$ is non-zero inside the set where the factor of the right-hand side of Equation 2.30 is non-zero. This factor is positive in the set

$$\widehat{C}_\alpha = \left\{ \boldsymbol{\xi}' = \sigma \boldsymbol{\theta} + \boldsymbol{\zeta} \in \mathbb{R}^3 \,\middle|\, |\sigma| < |\boldsymbol{\zeta}| \tan \alpha \right\} = \mathbb{R}^3 \setminus C'_{\pi/2 - \alpha}, \tag{2.31}$$

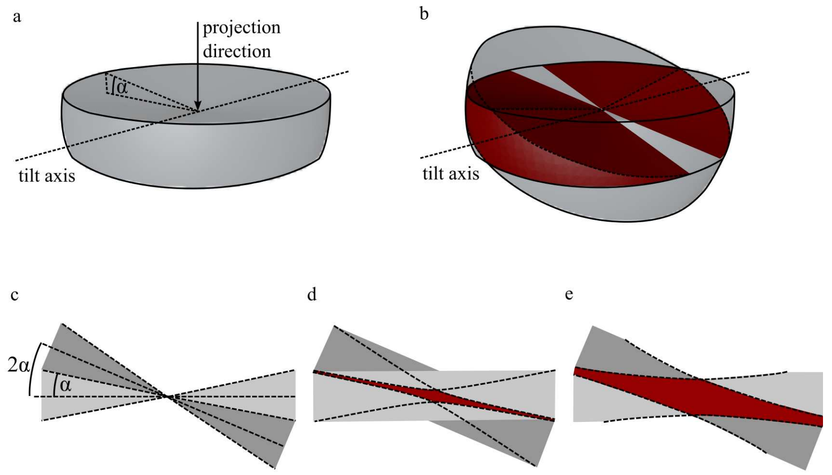

Figure 2.7: Geometric interpretation of the Fourier slice theorem for the STEM transform in Fourier space. a) The frequencies covered by one projection correspond to the shape of a double wedge of opening semi-angle α. b) If the tilt increment is chosen as $\Delta\beta = 2\alpha$, neighboring wedges overlap in a non-trivial shape (red). c) When considering a cross section through the origin and perpendicular to the tilt axis, the wedges seem to seamlessly cover the entire frequency space. d) A cross section shifted along the tilt axis reveals a complex-shaped region in frequency space that contains information from more than one tilt direction. e) A cross section even further along the tilt axis towards highest frequencies exposes that the region containing information from both tilt directions expands towards higher frequencies. Figure from (Dahmen, Kohr, et al., 2015).

with the double cone $C'_{\pi/2-\alpha}$ defined as in Equation 2.19 with $\pi/2-\alpha$ instead of α. This complementary double cone is formed by rotating a double wedge with opening semi-angle α around the symmetry axis $\boldsymbol{\theta}$ (Figure 2.7a). This result implies that $\widehat{f}$ and thus f can be recovered from a finite number of projections since $\mathbb{R}^3$ can be completely covered by a finite number of sets $\widehat{C}_\alpha$ with different directions $\boldsymbol{\theta}$, provided that data for all such directions is available, i.e. that the full angular range can be measured.

This situation can be interpreted as follows. If the tilt increment is chosen as $\Delta\beta = 2\alpha$, neighboring wedges touch in the orthogonal plane such that the entire frequency space is covered. In regions far from the symmetry center (lowest frequencies) of the double wedge, neighboring wedges overlap in a complex shape as shown in Figure 2.7b. This means that certain spatial frequencies are contained in more than one projection, which indicates that reconstructions may reveal an anisotropy of spatial resolution. However, as suggested by (Intaraprasonk et al., 2008), the double cone model can be in-accurate especially for high frequencies, a limitation which can be overcome by calculating the beam intensity distribution based on aberrations instead of merely approximating its shape. Another implication is that in the development of a direct inversion method for a CTFS, the weights of frequencies that are contained in more than one projection would need to be considered. These issues connected to spatial frequency distribution in the data go beyond the scope of this thesis but are worth investigating further in the future.

2.4.5 The Adjoint of the STEM Transform

As mentioned before, the back projection should be an operator that is the mathematical adjoint of the operator modeling the forward projection. In the case of square-integrable functions as applied in the current setting, the adjoint $\mathcal{A}^*_{\boldsymbol{\theta}}$ is defined by the relation

$$\int_{\mathbb{R}^3} [\mathcal{A}_{\boldsymbol{\theta}} f](\boldsymbol{v})\, g(\boldsymbol{v})\, \mathrm{d}\boldsymbol{v} = \int_{\mathbb{R}^3} f(\boldsymbol{u})\, [\mathcal{A}^*_{\boldsymbol{\theta}} g](\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} \tag{2.32}$$

which needs to be true for any two volumetric functions f and g. This relation corresponds to the scalar product identity in Equation 2.11 since the scalar product of two square-integrable functions f_1 and f_2 is given by the integral

$$\langle f_1, f_2 \rangle = \int_{\mathbb{R}^3} f_1(\boldsymbol{u})\, f_2(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u}. \tag{2.33}$$

The definition Equation 2.22 of the STEM transform is inserted into Equation 2.32. Using that $p_{\boldsymbol{\theta}}(\boldsymbol{v} - \boldsymbol{u}) = p_{\boldsymbol{\theta}}(\boldsymbol{u} - \boldsymbol{v})$, one can calculate

$$\begin{aligned}\int_{\mathbb{R}^3} [p_{\boldsymbol{\theta}} * f](\boldsymbol{v})\, g(\boldsymbol{v})\, \mathrm{d}\boldsymbol{v} &= \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} p_{\boldsymbol{\theta}}(\boldsymbol{v} - \boldsymbol{u}) f(\boldsymbol{u})\, g(\boldsymbol{v})\, \mathrm{d}\boldsymbol{u}\, \mathrm{d}\boldsymbol{v} \\ &= \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} p_{\boldsymbol{\theta}}(\boldsymbol{u} - \boldsymbol{v}) g(\boldsymbol{v})\, \mathrm{d}\boldsymbol{v}\, f(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} \\ &= \int_{\mathbb{R}^3} [p_{\boldsymbol{\theta}} * g](\boldsymbol{u})\, f(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u}, \end{aligned} \tag{2.34}$$

such that the adjoint can be identified as

$$\mathcal{A}^*_{\boldsymbol{\theta}} g = p_{\boldsymbol{\theta}} * g = \mathcal{A}_{\boldsymbol{\theta}} g. \tag{2.35}$$

In other words, the STEM transform is self-adjoint. In order to interpret this result, one must keep in mind that the STEM transform is defined in such a way that it maps a volume to a focal stack, not an individual image. In mathematical terms, both the searched-for tomogram f and one projection $g_{\boldsymbol{\theta}}$ are volumetric functions, i.e. functions $\mathbb{R}^3 \to \mathbb{R}$, and the back projection acts exactly like the forward transform, namely as a convolution with $p_{\boldsymbol{\theta}}$.

There is, however, a practical difference between forward and backward projections, which is explained in the following, starting with the case $\boldsymbol{\theta} = \boldsymbol{e}_z$. The adjoint $\mathcal{A}^*_{\boldsymbol{e}_z}$ applied to a data stack $g_{\boldsymbol{e}_z}$ can be rewritten as

$$\begin{aligned}
\mathcal{A}^*_{\boldsymbol{e}_z} g_{\boldsymbol{e}_z}(\boldsymbol{u}) &= \int_{\mathbb{R}^3} p(\boldsymbol{v})\, g_{\boldsymbol{e}_z}(\boldsymbol{u}-\boldsymbol{v})\,\mathrm{d}\boldsymbol{v} \\
&= \int_{C_\alpha} w(v_z)\, g_{\boldsymbol{e}_z}(\boldsymbol{u}-\boldsymbol{v})\,\mathrm{d}\boldsymbol{v} \\
&= \int_{\mathbb{R}} w(v_z) \int_{|\bar{\boldsymbol{v}}|<|v_z|\tan\alpha} g_{\boldsymbol{e}_z}(\boldsymbol{u}-\boldsymbol{v})\,\mathrm{d}\bar{\boldsymbol{v}}\,\mathrm{d}v_z \\
&= \int_{\mathbb{R}} w(v_z) \int_{|\bar{\boldsymbol{v}}|<|v_z|\tan\alpha} g_{\boldsymbol{e}_z}\big((\bar{\boldsymbol{u}}-\bar{\boldsymbol{v}}, z-v_z)\big)\,\mathrm{d}\bar{\boldsymbol{v}}\,\mathrm{d}v_z,
\end{aligned} \tag{2.36}$$

where 3D vectors are written as $\boldsymbol{u} = (\bar{\boldsymbol{u}}, z)$ and $\boldsymbol{v} = (\bar{\boldsymbol{v}}, v_z)$. The integral over C_α has been decomposed into an integral with respect to v_z along the z-axis and an integral over a disc with radius $|z|\tan\alpha$.

In practice, the projection $g_{\boldsymbol{e}_z}$ is not a continuous function but approximated by a stack of discrete 2D arrays, where each 3D point can be identified with an individual position of the focal spot. Thus, the local coordinate system of the projection stack consists of the beam direction $\boldsymbol{e}_z$ and two unit vectors in the xy-plane. Now, for one fixed value v_z in the outer integral in Equation 2.36, the inner 3D convolution integral

$$\int_{|\bar{\boldsymbol{v}}|<|v_z|\tan\alpha} g_{\boldsymbol{e}_z}\big((\bar{\boldsymbol{u}}-\bar{\boldsymbol{v}}, z-v_z)\big)\,\mathrm{d}\bar{\boldsymbol{v}} \tag{2.37}$$

is restricted to a single image in the stack since the symmetry axis of C_α corresponds to $\boldsymbol{e}_z$, one axis of the local coordinate system of the data. In other words, the domain of integration lies entirely in a single image from the stack, which corresponds to one focal length. This fact is advantageous from a computational perspective since this 3D convolution and the integration along $\boldsymbol{e}_z$ can be treated separately.

Furthermore, the same principle of corresponding axes is true for any beam direction $\boldsymbol{\theta}$, so the previously mentioned separation is possible for all focal series in the dataset. In this general case, the inner convolution integral reads as

$$\int_{\substack{\boldsymbol{\eta} \in \boldsymbol{\theta}^\perp \\ |\boldsymbol{\mu}| < |t| \tan \alpha}} g_{\boldsymbol{\theta}}\big((s-t)\boldsymbol{\theta} + (\boldsymbol{\eta} - \boldsymbol{\mu})\big) \, \mathrm{d}\boldsymbol{\mu} \tag{2.38}$$

for the decompositions $\boldsymbol{v}' = t\boldsymbol{\theta} + \boldsymbol{\mu}$ of the integration variable and $\boldsymbol{u}' = t\boldsymbol{\theta} + \boldsymbol{\eta}$ of the evaluation point. Written in coordinates, this integral can be expressed as

$$\int_{|\bar{\boldsymbol{v}}'| < |v_z'| \tan \alpha} g_{\boldsymbol{\theta}}(\bar{\boldsymbol{u}}' - \bar{\boldsymbol{v}}', z' - v_z') \, \mathrm{d}\bar{\boldsymbol{v}}' \tag{2.39}$$

This correspondence of axes does, however, not hold in the forward transform because the coordinate system of the tomogram f is fixed, which means that the 3D convolution integral is not restricted to one horizontal slice of the volume but involves several slices, dependent on the size and angle of the disc over which is integrated. Therefore, it can be expected that the forward projection is more expensive to compute numerically than the back projection, a behavior which clearly shows itself in the numerical tests (Section 3.4.1).

Conclusions on the Theory

The STEM transform was introduced in order to consider the convergent shape of the electron beam in aberration corrected STEM. The line model of the electron beam was replaced by a double-cone. The operator was investigated analytically and a number of properties are shown. The STEM transform is a generalization of the X-Ray transform for parallel illumination and contains the latter as the special case $\alpha \to 0$. Most notably, the STEM transform is self-adjoint, a theoretic result that can be exploited to implement efficient solutions to the tomographic reconstruction problem as detailled in the following chapter.

Chapter 3

Implementation and Evaluation

In this chapter, the CTFS is introduced as a new recording scheme for HAADF-STEM tomography. 3D information is acquired by mechanically tilting the specimen, and recording a through-focal series at each tilt direction. The tilt focal algebraic reconstruction technique (TF-ART) is introduced as a new algorithm to reconstruct tomograms from such CTFS. The feasibility of both, the image aquisition scheme and the reconstruction algorithm are demonstrated following the publications (Dahmen et al., 2014a; Dahmen, Kohr, et al., 2015) by presenting a workflow consisting of image aquisition, lateral alignment, axial alignment, and tomographic reconstruction (Figure 3.1).

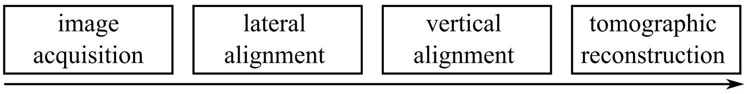

Figure 3.1: The entire workflow implemented for this thesis, consisting of image acquisition, lateral and axial alignment, and tomographic reconstruction.

3.1 Alignment

The first practical issue to be solved was that a TEM tilt-series needs to be aligned to correct for stage shifts occurring during tilting. This alignment is also required in pure tilt series recorded with parallel illumination and a large number of methods have been proposed for this problem (Section 1.2.1). However, these methods could not be directly applied to the case with a focal series for each tilt angle. It was observed that the alignment was sufficient within each focal stack but the image positions shifted as the tilt angle was changed. This observation is a consequence of the fact that the stage tilt is realized mechanically and thus subject to mechanical imprecisions while the focus change is realized via the magnetic lenses, i.e. does not involve any mechanical movements at all. The problem, however, was that occuring shifts did not only consist of a lateral ($x'y'$) shift but also of an axial shift. It was thus unknown how the indices of different frames of the focal series corresponded to the focus positions f, and it was not possible to find this relation with the existing algorithm. Therefore, two alignments were performed. First, the affine transforms for all images in the series were determined to bring the different projections into a common coordinate system, i.e. lateral alignment. Second, the parameters f_0 and Δf of the spatial positions of the focus planes needed to be found, i.e. axial alignment.

3.1.1 Lateral Alignment

For lateral alignment, the following procedure was used. The intensities for each pixel ($x'y'$) of the focal series were averaged, forming a vertical projection for the image stack of each focal series. The vertical projections were then combined into an image stack representing a conventional tilt-series. Next, the algorithm computed the affine transformations for the alignment using a standard method (Kremer, Mastronarde, & McIntosh, 1996). The determined transformations per tilt angle were applied to each image in that focal series. This method was possible because the focal series at each tilt angle did not contain noticeable lateral shifts in themselves. For cases with a significant lateral shift within a series, an additional alignment step could be added before the projection.

3.1.2 Axial Alignment

In the following, the procedure for axial alignment is described. The goal of the axial alignment was to find the relation between the index i of the image in the stack, the corresponding vertical focus positions of the first image f_0, and the focal distance between consecutive images Δf. This was achieved by searching first for nanoparticles in images corresponding to adjacent tilt directions (nanoparticle chain detection). Once the position of a certain nanoparticle was known in more than one image, its 3D position was estimated by means of triangulation. Finally, the algorithm detected which image in the stack was closest to focus for that nanoparticle and correlated the focus position to the 3D position of the nanoparticle, assuming that Δf is constant. The method is the first method for the axial alignment of STEM tilt series and as such a major contribution of this thesis. In the following, the method is described in detail.

3.1.3 Particle Chain Detection

A procedure for axial alignment was developed based on the automated identification of nanoparticles with high contrast in the images. First, the tilt-focal data was averaged into a tilt-series as described above. A high-pass filter with a 10 pixel cutoff was applied (ImageJ). After this, the stack was opened in TomoJ, using the normalization setting "Electron Tomo". Background removal was performed with a rolling ball radius of 15 pixel and smoothing enabled. Next, chains of objects were generated with a method described in detail elsewhere (Sorzano et al., 2009).

In short, nanoparticles were detected by searching for local maxima in an image. Those nanoparticles were then searched for in adjacent images in the tilt-series. The search was performed by predicting the nanoparticle position using affine transformations based on the tilt angle difference and consecutive local optimization of the correlation index. If the correlation index was greater than a given threshold, the regions in the two adjacent images were accepted to represent the same nanoparticle. A so-called particle-chain was thus generated from the local positions of the nanoparticle in consecutive images. A set of chains was generated aiming to track as many different nanoparticles as possible.

The particle-chain generation was performed in the software TomoJ using the following settings: algorithm = "critical points – local maxima", number of seeds = 20, number of best points to keep in each image = 40, length of landmark chain = 11, patch size in pixel = 14, minima neighborhood radius = 8, fiducial markers = yes. The found nanoparticle positions in the images were exported for further processing.

3.1.4 Nanoparticle Position Triangulation

Next, the algorithm estimated the 3D position of an individual nanoparticle by triangulation. It selected two images in which the 2D positions of the same nanoparticle were known. Based on both known lateral positions and the tilt angle, 3D lines were created by assigning an origin and a direction. The line origin was the nanoparticle position in the image plane while the line direction was the tilt direction associated with that image, assuming parallel rays.

In the ideal case of perfect alignment, the two lines corresponding to the same nanoparticle would intersect in the center of that nanoparticle. In practice, the lines did not intersect due to alignment errors. Therefore, the line segment of closest distance between them was calculated. The midpoint of this segment provided an estimate of the 3D position of the nanoparticle, relative to the tilt axis. The length of the segment was considered as a measure of the alignment error. The above procedure was repeated for all possible line-pairs of the same particle-chain. The midpoints were then averaged to obtain the most precise position estimate. The procedure was applied to all particle-chains, each giving the 3D position of a different nanoparticle.

3.1.5 Estimation of Focal Values

Our algorithm computed the parameters for the axial alignment. For a given tilt direction, the software considered all 1,516 determined nanoparticle center positions using the 2D ($x'y'$) position of the nanoparticle from the particle-chain. For each position, it searched all frames in the focal series, applying a low-pass-filter with a radius of 2 pixels in the lateral direction to suppress noise. The image i with the highest intensity at the pixel $x'y'$ was defined as the best focus for this nanoparticle. The corresponding index was called i_{focus} (Figure 3.2a).

This procedure was repeated for all known nanoparticles and the algorithm created (in principle) a plot displaying the axial distance of the nanoparticle to the tilt axis z' against i_{focus} (Figure 3.2b). Hereby, z' is obtained by rotating the 3D position xyz of the particle with respect to the current tilt direction. A linear trend was fitted through the plot using linear least squares regression. Since the data was recorded with identical focus steps between the images in a focal series, the slope of the trend gave the relative distance between consecutive focus positions Δf, while the intersection with the f-axis gave the focus position of the first image with respect to the tilt axis f_0. The operation was repeated per tilt direction, so the software computed one value for f_0 and Δf for every image stack.

3.2 Tomographic Reconstruction

In the following section, tilt- focal algebraic reconstruction technique (TF-ART) is presented, a new method of volume reconstruction applicable

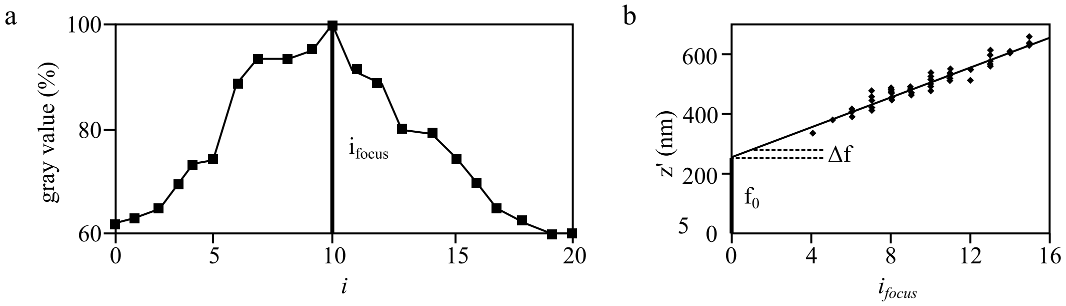

Figure 3.2: Estimation of focal positions for axial alignment. a) Plot of the pixel intensity of an individual nanoparticle versus the focus index i as a measure of the focus index where the nanoparticle was best in focus i_{focus}. b) Plot of the vertical position z' versus i_{focus} for all nanoparticles. A fitted linear trend is also shown. The slope of this trend equals Δf. The intersection with the f-axis gave the location of the first image plane f_0. Figure from (Dahmen et al., 2014a).

to a combined tilt- and focal series (CTFS). TF-ART is a generalization of the block iterative algorithm family (Censor, 1990) and is defined by a forward projection that implements the double cone model of the electron beam, an (unmatched) back projection based on a heuristic weighting factor and a special update loop as explained below. Figure 3.3b depicts the algorithm as a block diagram.

After the presentation of TF-ART, an alternative, matched backprojection operator is presented, which is the adjoint of the forward projection. Using the matched operator and changing from the block-iterative update scheme to a SART-type algorithm, it a second variation of the algorithm is presented. Contrary to TF-ART, the algorithm based on the matched backprojeciton is an instance of Kaczmarz algorithm family.

3.2.1 Forward Projection

The forward projection operator (STEM transform) can be expressed as a convolution, i.e. the value of a point in the projection is the convolution of the volume function and function modelling the probe shape. Thus, the computational task for the forward projection is to compute a weighted integral of the volume function, where the weighting factors correspond to the factors of the double cone in the STEM transfrom (Section 2.4, Equation 2.23).

In order to compute this volume integral, the algorithm used a Monte-Carlo technique, approximating the volume integral with multiple individual line integrals. The lines were chosen such that they formed a specific sampling

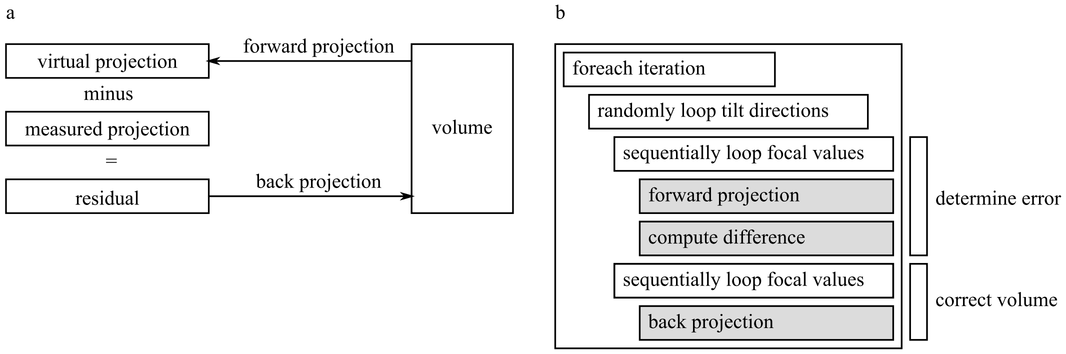

Figure 3.3: The overall reconstruction algorithm. a) Visualization of the dataflow. Starting from an initial reconstruction volume, residuals are generated by a forward projection and per-pixel subtraction from a measured projection. The volume is then corrected using a back projection operator. b) The nesting of the reconstruction loops. Operations that are implemented as kernels on the GPU are marked as light gray. All virtual projections from one direction are generated before a correction is applied to the volume. Figure from (Dahmen et al., 2014a).

pattern as described below. Each line integral was computed using GPU-based volume ray-casting (Engel, Hadwiger, Kniss, Rezk-Salama, & Weiskopf, 2006; Rodriguez et al., 2013). The integrals of the individual lines were then averaged to estimate the volume integral of the double cone. Figure 3.4a depicts this principle. The sampling method was inspired by the way the focal depth of optical camera systems is simulated using Whitted-Style ray tracing (Whitted, 1980).

An alternative approach would have been to implement the integration using a modified 3D Bresenham (Bresenham, 1965) algorithm for conservative line drawing. Conservative in this context means that the algorithm iterates over all voxels that have a distance to the line below a given threshold, thereby extending the line to a cylinder. In computer graphics, the Bresenham algorithm is typically used to draw lines with a given thickness. By replacing the fixed radius with a radius that is a linear function of the ray distance, the algorithm can be modified to iterate over double cones instead of cylinders. The Bresenham implementation has the advantage of avoiding sampling errors and can be more efficient than the Monte-Carlo technique if a very high precision of the integral computation is needed. However, the method has the drawback that it iterates the voxel contributing to the double-cone, but as such does not present a way to compute the weights inside the voxel. The weights are typically assumed to be constant within the voxels, which makes the computation trivial but can be problematic in the important area close to

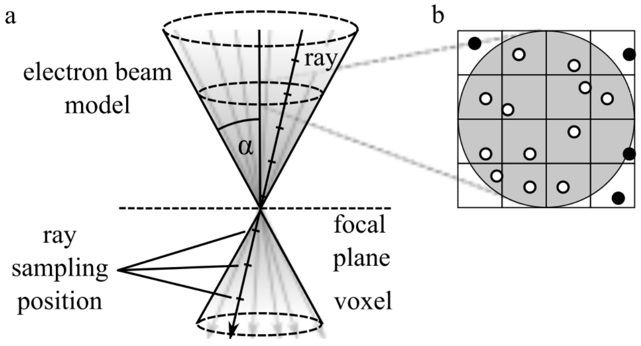

Figure 3.4: Implementation of the forward projection operator. a) The electron beam is modeled as a double cone. The intensity of a pixel is computed by integrating the gray-values of the volume inside the double cone. Hereby, it is approximated using several rays, which are then integrated using ray-casting and averaged. b) In the Stratified Sampling scheme, a grid is placed over the circle which represents a horizontal slice through the volume. In every cell of the grid, one sample (white dot) is placed at pseudo-random position. Samples outside the disc are rejected (black dots). Figure from (Dahmen et al., 2014a).

the focus point of the double-cone. Ironically, the straight-forward solution to this problem is to better approximate the weights using stochastic sampling, thereby re-introducing all issues of the Monte-Carlo technique.

Stratified Rejection Sampling

As a method to randomly place the individual lines in the double cone while maintaining roughly uniform sampling over the domain, "stratified sampling" (Cook, 1986) was used. Each line was specified by two points. The first point was the focus point of the double cone. In order to specify the second point a horizontal cut through the cone was considered specifying a circular disc. A 2D grid was placed over this disc and within each grid cell one point was placed at a pseudo-random location. If the point happened to be outside the disc, the sample was rejected and the corresponding line was not considered during integration. The method is called "rejection sampling" (Robert & Casella, 2005) and is frequently used together with ray-casting. Figure 3.4 depicts the principle.

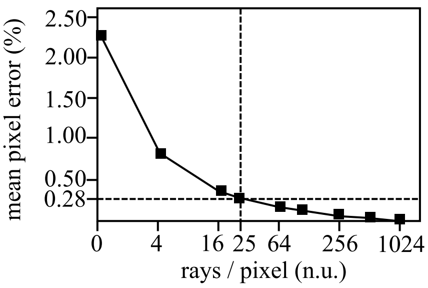

Figure 3.5: The per-pixel-error of a forward projection plotted over the number of rays per pixel that was used to approximate the double cone. Figure from (Dahmen et al., 2014a).

In summary, a forward projected image was generated by stepping through all pixels of the projection. For every pixel, the probe was approximated by a double cone according to known tilt angle and focal position. Inside this double cone, the volume integral of the voxel gray values, weighted by the local current density of the beam (because rays converged), was computed. The computation was performed by a cone tracing implementation based on ray casting and stratified rejection sampling. The resulting projection corresponds to a single slice of the focal series, assuming a simplified PSF and the absence of aberrations, alignment issues, noise, and drift.

Influence of Sample Count on the Precision of the Forward Projection

The implementation of the STEM transform used in the forward projection relied on stochastic ray tracing to compute the volume integral in the double cone. In order to determine the required number of rays per pixel for the forward projection, it was tested experimentally how quickly the computational error dropped as the number of rays used to approximate the double cone was increased (convergence rate). Figure 3.5 shows a plot of the mean pixel error as percentage of the maximum intensity over the number of rays per pixel.

In order to determine the per pixel error, a ground truth image was generated using a very high sample count (100.000 samples per pixel). It was confirmed that at this high sampling, adding further samples did lead to no difference in the result within computation precision and the ground truth image was used as reference to measure the sampling error of the forward projection. In the case that 25 rays per pixel were used to approximate the double cone, the remaining mean error was 0.28% of the maximum intensity. This pixel error had no measurable impact on the overall reconstruction quality so 25 rays per pixel were thus used for all reconstructions.

3.2.2 Unmatched Back Projection Based on Heuristic Weighting Factor

The volume reconstruction algorithm also needs a back projection. For the first proof-of-concept study (Dahmen et al., 2014a), it was not possible to use the matched back projection, because the adjoint of the projection was initially not known. Instead, the effect of the PSF was corrected with a heuristic weighting method. The method is based on the idea that an individual projection image should only influence that part of the volume where the image is best in focus (Figure 3.6).

The back projection described in the following section corrects the volume for an individual residual image, corresponding to one tilt direction and one focal value. This is different from the definition of the STEM transform operator (Section 2.4.2), where one application of the transform corresponds to the variation of the focus over the entire volume, i.e. one stack of images. In order to correct the volume for one tilt direction (one full application of the operator), the back projection was executed consecutively for every focal value.

An individual application of the back projection was implemented by looping over all voxels in the relevant reconstruction volume. The center of each voxel was projected to the image plane of the projection by multiplication with the 4x4 matrix representing the parallel projection corresponding to the tilt angle. Bilinear interpolation was used to look up the residual value at the projected pixel position. The voxel was then corrected with the residual value, modified by a weighting factor described below.

This method is functionally equivalent to looping over the residuals pixel-by-pixel and projecting the pixel value to the reconstruction volume using ray-casting. However, the voxel-by-voxel approach maps better to current GPU hardware because it avoids scattered memory-write operations and synchronization issues and thus results in higher performance. A more detailed discussion of the topic is given elsewhere (W. Xu et al., 2010).

The back projection was limited to that part of the volume where the corresponding projection contains the most information, i.e. information that has not been blurred by integrating one cone. We achieved this by introducing a heuristic weighting factor Γ.

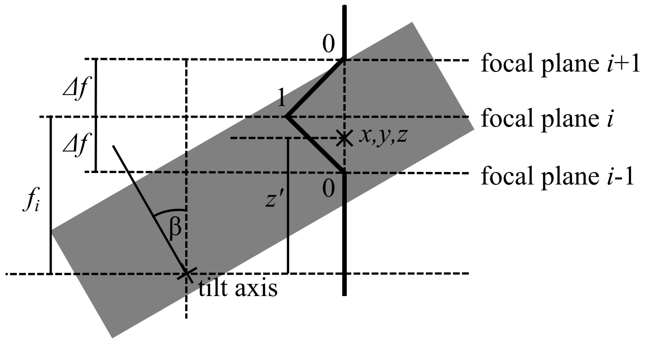

Figure 3.6: The weighting factor used in the back projection operator. The regularization factor $\Gamma(u)$ is expressed as a function of the position u in the volume. $\Gamma(u)$ equals one exactly inside the focal plane and drops linearly to the previous and next focal plane. Figure from (Dahmen, Kohr, et al., 2015).

$$\Gamma(u,\beta) := \lambda \cdot \begin{cases} 0 & \text{if} \|z' - f_i\| > \Delta f \\ 1 - \frac{\|z'-f_i\|}{\Delta f} & \text{else} \end{cases} \tag{3.1}$$

Here, $\Gamma(u,\beta)$ is a function of the position $\boldsymbol{u} := (x,y,z)$ of the center of the voxel that is currently being corrected, and the tilt direction. z' is the perpendicular distance from the lateral plane containing the point u to the tilt axis and given by $z' = \cos\beta z - \sin\beta x$, so z' depends on the tilt direction. λ is the relaxation parameter typically used in algebraic reconstruction methods (Gordon et al., 1970). $\lambda = 0.3$ is used for all experiments.

The idea behind the formula for $\Gamma(u,\beta)$ is that information from a focal plane should only influence the region of the volume close to the focal plane. Between the individual planes, linear interpolation is used to achieve a smooth transition. Figure 3.6 is a schematic representation of this weighting scheme. $z' - f_i$ is the axial distance from a voxel center to the focal plane i. The regularization factor Γ is zero everywhere except in a slice of thickness Δf on both sides of the current focal plane, so an individual application of the back projection operator only corrects a slice of thickness $2\Delta f$, the remainder of the volume remains unchanged. In order to correct all voxels in the volume, the back projection is executed once for every residual image from one tilt direction as determined by the reconstruction loop. After the execution of a tilt direction, every voxel in the volume was changed twice, once for each of the two focal planes closest to the voxel, resulting in a linear interpolation between the two relevant residual values for each voxel.

3.2.3 Matched Back Projection Based on Adjoint Transform

The unmatched back projection described in the previous section results in good reconstruction results, but showed slow convergence behavior. In a second study (Dahmen, Kohr, et al., 2015), the correctly matched adjoint of the STEM transform was therefore derived analytically and evaluated experimentally. A detailed derivation of the adjoint has been given in Section 2.4.5 and results in the representation:

$$\mathcal{A}_\theta^* h_2 = p_\theta \ast h_2 = \mathcal{A}_\theta h_2. \tag{3.2}$$

In other words, the STEM transform is self-adjoint. With this results on the adjoint of the STEM transform, it is possible to implement the matched back projection corresponding to the forward model. In the following, details on a software implementation of this operator are given. The implementation exploits the fact that the STEM transform can be written as a linear convolution of the images and the double cone that models the electron beam (Figure 3.7). In the following sections, we describe how a combination of prefiltered images and linear interpolation is used to achieve an efficient implementation of the operator.

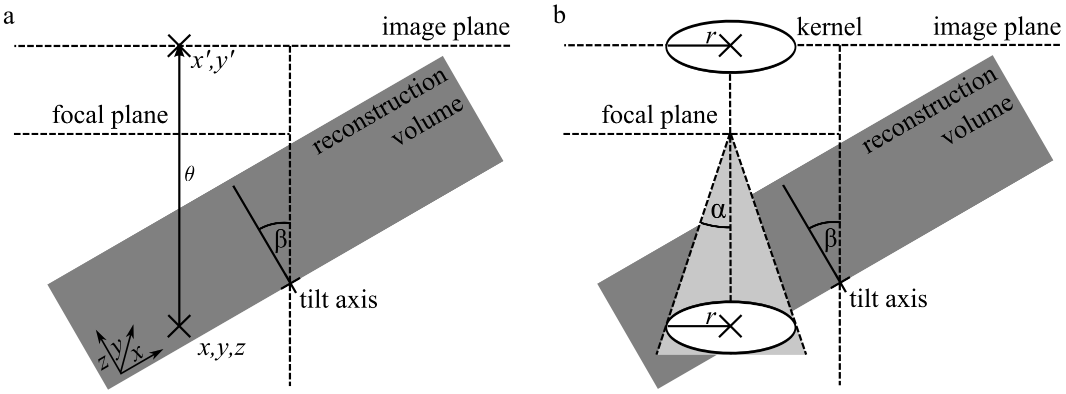

Figure 3.7: Geometry of the back projection. a) The center of each voxel (x, y, z) is projected along θ on the image plane at pixel coordinates (x', y'). b) The residual is convolved with a separate kernel for each voxel. The kernel corresponds to a lateral cut through the electron beam, at the position of the voxel. Figure from (Dahmen, Kohr, et al., 2015).

One application of the back projection, corresponding to the correction with respect to an individual slice of the projections, i.e. one value of z', is implemented by a parallel loop over all voxels in the reconstruction volume. As for the unmatched back projection, the implementation operates on individual slices, not full 3D projections as discussed in Section 3.2.4. The gray value of each voxel is corrected by the addition of a correction term, computed as follows. The center of each voxel is rotated according to the current tilt direction. The correction term is equal to the value of the convolved residual image at position $\boldsymbol{u'} = (x', y')$, corresponding to the coordinates of the rotated voxel center (Figure 3.7a). The value of this pixel in the residual image resulted from the convolution with a specific kernel that has the shape of a perpendicular cut through the double cone C_α of the beam model (i.e. a circle) at the distance $z' - v'_z$ of the voxel being corrected from the current focal plane (Figure 3.7b). Thus, the convolution kernel is different for every voxel and given by

$$I(x', y') := \begin{cases} 0 & \text{if } \sqrt{x'^2 + y'^2} > r \\ \dfrac{1}{\pi r^2} & \text{else,} \end{cases} \tag{3.3}$$

where $r = |z' - v'_z| \tan \alpha$. It has intensity zero outside the circle by definition. Inside the circle, the intensity was chosen according to the normalization criterion, such that the pixel values did add up to one and the total intensity of the residual was preserved after the convolution.

The radius of this convolution kernel can be as much as 43 pixels in the areas of the volume farthest from the focal plane, assuming a volume thickness of 1024 voxels and a beam opening angle of $\alpha = 42$ *mrad*. Computing the convolution for every voxel separately is therefore prohibitively slow. In order to overcome this problem, the residual image is pre-filtered for a suitable set of radius values $r \in \{r_0, \dots, r_n\}$ as explained below and stored in memory. Linear interpolation is then used to approximate values between the two nearest pre-filtering values of r_i.

In order to compute the pre-filtered residual images, the residual is first transfered to Fourier space by means of a 2D Fast Fourier transform (FFT). The library clAmdFFT (AMD, 2013) is used to achieve this efficiently on the GPU. The filter kernel is then generated for each radius r_i and also transfered to Fourier space. The convolution is computed by means of a complex multiplication, and the result is transfered back to real space by means of an inverse FFT, again using the same library. In order to compute the pre-filtered residual at N different radius values, the algorithm needs to perform $2N + 1$ 2D FFT computations. Computational cost for the generation of the filter kernel and the complex multiply can be neglected in comparison to the cost of the FFT, and the resulting filtering step shows feasible performance.

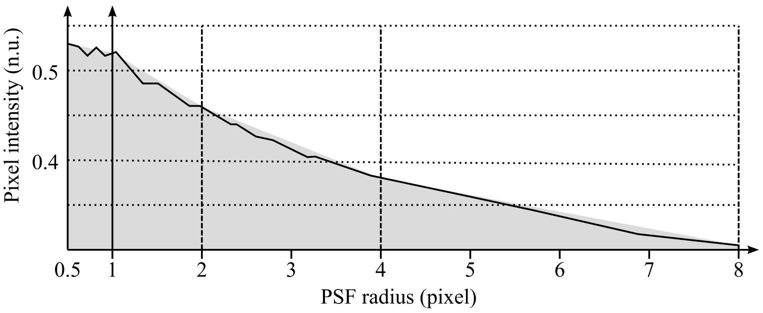

Figure 3.8: The intensity of the center of an arbitrary nanoparticle as a function of the prefilter radius. The function is sampled at positions of exponential distance (vertical lines) and linear interpolation is used between the sample positions. The linear approximation of the function is shown in gray. Figure from (Dahmen, Kohr, et al., 2015).

Sampling positions $\{r_0, \ldots, r_n\}$ are chosen such that the error introduced by the interpolation has little influence on the reconstruction results. The radii were determined experimentally by computing the pre-filtered residuals for one residual image in very small steps ($\Delta r = 0.1$). The center of one nanoparticle was selected, and the intensity of this pixel was plotted as a function of the prefilter radius r, as shown in Figure 3.8. As can be seen, the intensity exhibits relatively rapid changes for small radius, but shows almost linear behavior for larger values of r. The sampling positions $\{r_0, \ldots, r_n\}$ were therefore chosen using an exponential pattern $r_i := 2^{(i-1)}$. This way, the first sample is generated for $r = 0.5$, which corresponds to a filter kernel consisting of a single pixel, i.e. no blur. The largest required radius was $r = 64$, so a total of 8 images had to be calculated.

3.2.4 Update Loop

The TF-ART algorithm loops over all tilt directions in pseudo random order. For each direction, it steps through all focus positions sequentially. The residual images for all focus positions of one tilt direction are computed and stored on a stack. Then, the volume is corrected for all residuals of this direction by iterative execution of the back projection before moving to the next direction. In this scheme, one iteration of the algorithm refers to processing the images from all directions and all focus values once. One update block (in the sense of block iterative algorithms (Censor, 1990)) refers to the execution of the forward and back projection for all images from one direction.

This scheme is chosen because for technical reasons the implementation of the back projection corresponds to the processing of an individual image (i.e. one focus value), while the theoretic definition of the back projection operator considers the processing of an entire image stack (i.e. one variation of the focus over all possible values). By sequentially processing all images of one direction inside one update block, we achieve that one update block equals one execution of the STEM transform operator. The back projection for an individual image only influences a slice of thickness Δf around the focal plane, so every voxel is corrected twice (once each for the two images with their focal planes closest to the voxel). The weighting factors of the two correction sum to one such that the desired amount of correction for the back projection operator is achieved.

The matched back projection, however, does not feature a heuristic weighting factor, so every voxel is corrected by every residual with weight Λ as in Equation 2.13. In order to avoid over-correction that would lead to oscillation and ultimately divergence, one could normalize the correction weight by the block size (i.e. correct only by Λ/n, where n is the number of focal lengths per direction). This method works but leads to very slow convergence rates. In the case of the matched back projection, the block-iterative scheme was therefore abandoned when working with the matched back projection in favor of a SART- type algorithm where the residual for only one focal length is computed before the back projection is executed for this residual image.

3.3 Evaluation and Results

The concept CTFS and both reconstruction algorithms were evaluated by applying the method to an exemplary sample. The study consisted of the preparation of a sample, aquisition of the images, alignment, tomographic reconstruction, and analysis of the reachable resolution, particularly of the effect of the missing wedge on the axial elongation factor. The sample was chosen primarily by the specimen thickness. In order to demonstrate the effect of the CTFS and particularly the advantageous properties of HAADF STEM imaging, a sample was selected that was too thick for reasonable imaging in CTEM. On the other hand, the thickness was limited to a range where beam blurring due to multipe scattering does not yet have relevant impact on the beam shape. Those two conditions restrict the sample thickness to about 500 nm to 2 μm, so a whole human cell was choosen as a sample. The staining with gold markers was required to achieve sufficient contrast in the HAADF imaging mode. Other than this, the sample had no relevance for this study and was chosen because it was available from previous work (Baudoin, Jinschek, Boothroyd, Dunin-Borkowski, & de Jonge, 2013).

3.3.1 Sample Preparation

Macrophages derived from monocytes (THP-1 cells, American Type Culture Collection) were grown directly on electron transparent silicon nitride TEM windows supported by silicon microchips (Ring, Peckys, Dukes, Baudoin, & de Jonge, 2011). The growth occurred in phorbol-12-myristate-13-acetate supplemented medium (Jerome, Cox, Griffin, & Ullery, 2008). Native low-density lipoprotein (LDL) was conjugated to $16 \pm 3\,nm$ or $7 \pm 1\,nm$ gold nanoparticles. Cell samples were incubated with $16\,nm$ LDL-gold for the first day and with $7\,nm$ LDL-gold for the second day (Baudoin, Jerome, et al., 2013). The incubation took place at $37°\ C$ in 1% fetal bovine serum medium with an equivalent concentration of $8\,\mu g/mL$ LDL. To prepare the samples for electron microscopy, the cells were rinsed with phosphate buffer saline, fixed with glutaraldehyde 2.5% in 0.1 molar sodium cacodylate buffer / 0.05% CaCl2, post-fixed with an ultra-low concentration (0.001%) of osmium tetroxide, gradually dehydrated with ethanol, and finally critical point dried with liquid carbon dioxide. To increase resistance to electron beam damage, a layer of about $20\,nm$ of carbon was evaporated on the samples (Dukes et al., 2011). The carbon was applied using an electron beam evaporator with a base pressure of $5 \pm\ 10 - 7\,torr$ for 45 minutes. Additional staining with e.g. lead was avoided to be able to image through the entire cell.

3.3.2 Data Acquisition

STEM images were recorded at $300\,kV$ with a transmission electron microscope equipped with a probe corrector for spherical aberrations (Titan 80-300, FEI, Hillsboro, OR, USA). Images were acquired at $160,000\times$ magnification. The objective aperture semi-angle was $49.1\,mrad$. The focal-series consisted of 41 images separated by $50\,nm$ in axial direction, of which the 20 images in the middle vertical range were selected for further processing as the first and last images were entirely out of focus. Images of 512×512 pixels were recorded with an acquisition time of $12\,\mu s$ per pixel. The tilt-series were recorded with specimen tilts ranging from ±40° at 5° increments (higher tilt angles were not possible because the edges of the microchip masked the cell at higher tilt angles). A script (written in Java) controlled the FEI microscope for an automated acquisition of the focal series. After changing the tilt angle, the region of interest was realigned, and the probe was refocused. Figure 3.9 shows two images of the exemplary specimen. The fact that for every tilt direction about 40 images were recorded implies an increase in the overall electron dose. However, this increase is partially compensated by the fact that a larger tilt

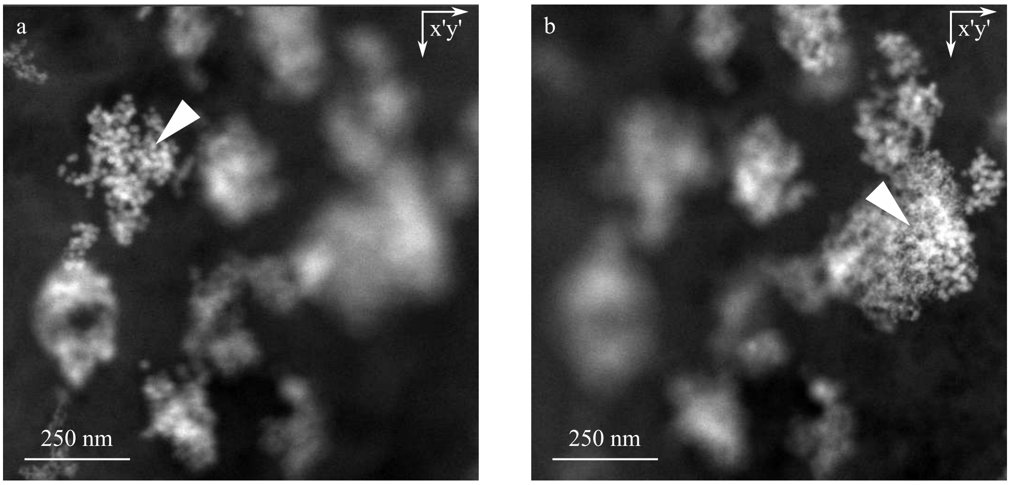

Figure 3.9: Projections of the original CTFS high annular dark field STEM data (input). The sample was a whole mount macrophage cell containing gold nanoparticles of two different sizes distributed in clusters throughout its volume. a) and b) Two different images of a CTFS. Both show nanoparticles at the tilt direction -40° at different focal positions. The white arrows indicate the sections of the images that are in focus. Figure from (Dahmen et al., 2014a).

increment can be used compared to a conventional tilt-series. It would also be possible to record a fewer number of images per focal series than acquired in this study, which should be further investigated.

3.3.3 Evaluation of Alignment

The CTFS was aligned in lateral direction using a standard method to a precision of ≈ 1 pixel. Hereafter, the alignment in axial direction was performed using the newly developed procedure involving particle chains (Section 3.1). In total, 134 particle-chains were generated, containing a total of 1,516 known 2D particle positions. For every tilt direction, about 160 (standard deviation $s = 14$) different 2D nanoparticle positions were known, which allowed the procedure later to reach the required precision in axial alignment.

The 2D nanoparticle positions were used to determine the focal parameters f_0 and Δf by fitting a linear trend (Subsection 3.1.5). Hereby, Δf was considered constant across tilt directions. However, the fitting of the linear trend did not reach the same measure of confidence for all tilt directions on account of the difference in vertical separation of the nanoparticles, i.e. at higher tilt angles, the nanoparticles as seen from the electron beam were spread out more than at lower tilt angles, leading to a more accurate estimate of Δf. Thus, for the purpose of estimating Δf, only the 11 directions with the highest measure of confidence were considered, which were the directions with the highest tilt angles. By combining those values, a mean value of $\Delta f = 60\,nm (s = 3\,nm)$ was calculated. This value differs significantly ($p < 6.6 \cdot 10^{-6}$) from the nominal value of $\Delta f = 50\,nm$ that was expected from the microscope control. The reasons for this discrepancy are likely specific to the exact type of microscope. The computed value for Δf was used in all reconstructions.

In addition to the focal distance, the value of f_0 had to be calculated independently per tilt direction. The precision of the measurement of f_0 was quantified by calculating s of the constant coefficient of the linear fit using statistical standard methods. For f_0, a standard deviation of $s = 18\,nm$ was calculated, which corresponds to a precision of $\pm 37\,nm$ assuming a 95% confidence level. Because the electron beam had an opening semi-angle of $a = 41\,mrad$, this axial alignment error corresponded to a blurring with a radius of $1.5\,nm$ (0.7 pixel) which was considered to have no observable impact on the reconstruction quality. The alignment procedure relies on the presence of gold nanoparticles but fiducial gold markers, commonly used for STEM or TEM tomography (Lawrence, 1992), should work as well.

3.3.4 Evaluation of Unmatched Back Projection

In order to evaluate the TF-ART reconstruction, we compared the results with a conventional tomographic SART reconstruction of a STEM tilt-series of a similar sample (Baudoin, Jerome, et al., 2013) containing nanoparticles of similar but slightly smaller sizes (5 and $14\,nm$) and recorded using a smaller pixel size of $0.67\,nm$ than the $2.3\,nm$ used here.

When comparing a CTFS to a conventional tilt-series, the choice of the beam convergence semi-angle is a crucial issue. STEM tilt-series are best recorded with a very small beam opening semi-angle such as $2\,mrad$ as proposed in (Hohmann-Marriott et al., 2009). This is the case because the limited DOF resulting from a larger beam opening semi-angle cannot be compensated and would lead to geometric blurring. On the other hand, in a CTFS, information from different focal planes is used to enhance the tomogram. Consequently, a large beam opening semi-angle is favorable. Tomograms were thus compared that were recorded under conditions chosen for the respective method, i.e. $2\,mrad$ were used for the tilt-series and compared it to a CTFS recorded with $41\,mrad$ ($\approx 2.4°$) beam opening semi-angle. For the comparison, a tilt-series was recorded of a dataset of whole cells containing LDL coated

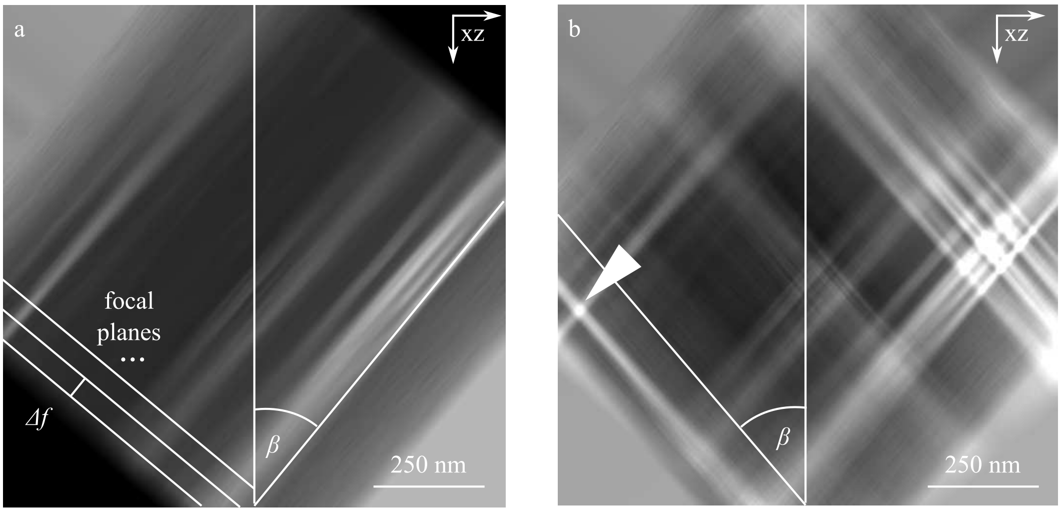

Figure 3.10: Intermediate results of the back projection operator. a) Intermediate tomogram after all residuals from one tilt direction were processed. The focal planes are displayed as white lines. b) Intermediate tomogram after the processing of a second direction. The white arrow marks the location of a nano-particle. Figure from (Dahmen et al., 2014a).

gold nanoparticles (Baudoin, Jerome, et al., 2013). The sample was imaged with STEM with a tilt range of 76° (−38° to +38°) in 4° tilt increments and reconstructed using SART. This reference tomogram is shown in Figure 3.11b.

Figure 3.10a shows the intermediate volume after all images of one tilt direction were processed. After only one direction, nanoparticles were reconstructed as elongated streaks in the tilt direction. The focal planes were perpendicular to the tilt direction and are shown as white lines in the image. Information exactly in one focal plane was identical to the image in the input stack corresponding to this plane. Between the focal planes, linear interpolation was used. After two iterations had been processed (Figure 3.10b), the positions of the nanoparticles started to show as intersections of the streaks.

The reconstruction was performed using 120 iterations of TF-ART. The resulting tomogram is shown in Figure 3.11a, the white arrow marks the position of the same nanoparticle in all three directions. The gold nanoparticles are clearly visible against the background, and their spherical shape is reconstructed accurately considering the limited tilt range, i.e. as somewhat oval shapes as seen from yz or xz projection. Examination of the xz slice shows some additional artifacts, especially a star-like structure in the direction of the maximum tilt directions. Figure 3.12 shows a perspective rendering of the tomogram, with a contrast transfer function applied for coloration.

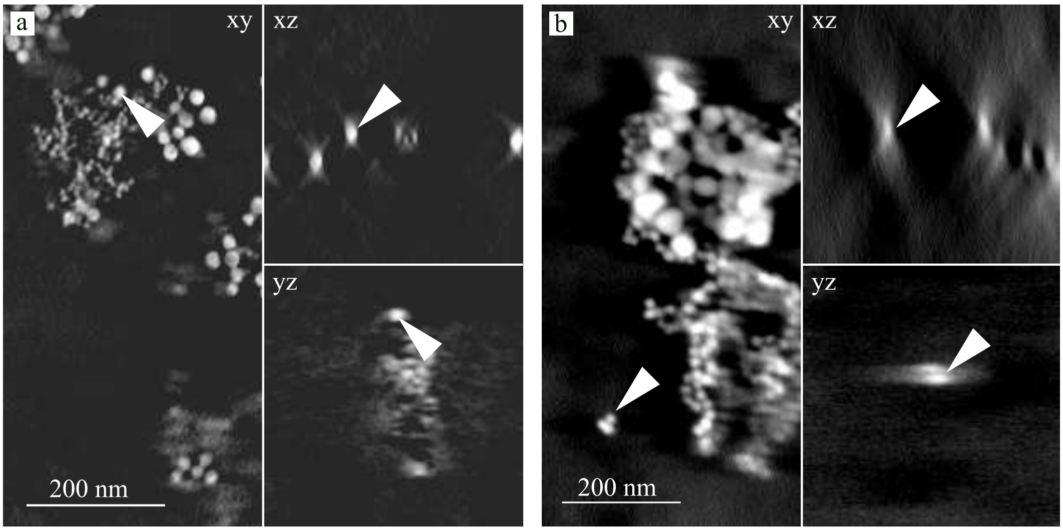

Figure 3.11: Comparison of the results of the reconstructions of the combined tilt-focal series STEM data with the reconstruction of a STEM tilt series. a) Tomogram of the TF-ART reconstruction from a combined tilt-focus series. b) Tomogram of the SART reconstruction from a STEM tilt series of a similar sample. The white arrows mark the same nanoparticle in each of the views. Figure from (Dahmen et al., 2014a).

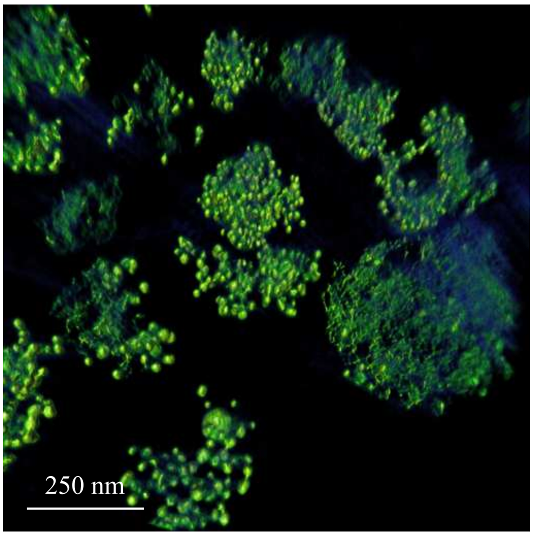

Figure 3.12: Perspective rendering of the tomogram of the TF-ART reconstruction with contrast transfer function for coloration. Figure from (Dahmen et al., 2014a).

Axial Resolution

One main issue with tilt-series based tomograms that was improved is axial elongation, i.e. the effect that tomograms have a lower resolution in the axial direction than in the lateral direction. This effect is typically quantified by the axial elongation factor (e_{yz}) (Section 2.1.5). For the tilt-series, a value of $e_{yz} = 2.8 \pm 0.5$ was measured. For the CTFS, measurement gave $e_{yz} = 2.0 \pm 0.5$. This corresponds to an improvement of about 29% and is statistically clearly significant ($p < 7.0 \cdot 10^{-4}$). The result constitutes a major improvement on one of the two most relevant restrictions in electron tomography compared to a tilt series reconstruction. The measurement for a tilt-series differs from literature which reports $e_{yz} = 2.5$ (Baudoin, Jerome, et al., 2013). This difference could be attributed to the fact that in the literature one of the smallest particles was selected manually while here we selected a total of 17 nanoparticles randomly for computing the mean. However, even when comparing to the value reported in the literature, this combined tilt- and focal series still results in an improved axial elongation of 20%.

Directional Dependence of Elongation Factor

After the very promising results on the axial elongation factor, we wanted to exclude the possibility that the tomogram exhibits lowest resolution in a direction other than the axial direction as a result of the limited number of tilt directions. The angle-dependent elongation factor $e_{x\gamma}$ was measured as explained in Section 2.1.5. A plot of $e_{x\gamma}$ over γ is shown in Figure 3.13. It can be seen that the tilt-series has a larger elongation factor than the CTFS around the axial direction (90°), which is in line with the earlier measurements of axial elongation factor. The elongation factor was up to 1.27 times larger for a CTFS between 110° and 135° direction. This effect was present only on one side (the opposite range of 45° to 80° did not suffer from this effect), so it is presumably the result of an alignment error. We concluded that as expected, the tomogram has lowest resolution in axial direction as a result of the missing wedge, but the effect is better than in the case of a tilt series.

Frequency Domain

To further evaluate the information obtained from the CTFS, two xz slices of the tomograms of the CTFS, and of the tilt series were transferred into the frequency domain (Fourier transform). In the case of the tilt series (Figure 3.14a), sharp streaks are present corresponding to the tilt directions. The

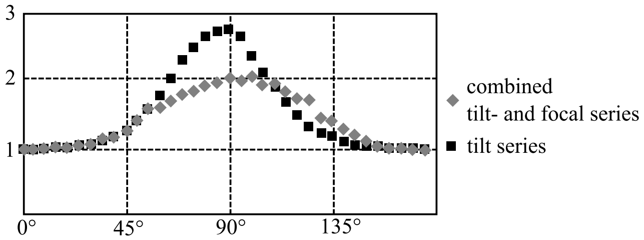

Figure 3.13: Analysis of the gain of information in vertical direction. Plot shows the direction dependent elongation factor ($e_{x\gamma}$) over the angle to the lateral plane, compared between the CTFS, and the tilt series. Figure from (Dahmen et al., 2014a).

"missing wedge" effect is also clearly visible, and the vertical direction contains hardly any signal components. The information obtained in the vertical direction is thus very limited.

In the case of the CTFS (Figure 3.14b), the streaks corresponding to the tilt directions are less pronounced and spread over an angular region equal to the beam-opening angle, so additional information is present between the tilt directions. Note that the streaks in the tilt-series would be less pronounced for a dataset containing more tilt planes with a smaller tilt angle spacing. The missing wedge is still visible for the combined tilt- and focal series but in the central vertical region low spatial frequency signal components are now present (white arrow). Thus, the CTFS results in additional information in the axial direction compared to a pure tilt series, which can be observed in real space as a reduction of the axial elongation.

3.3.5 Evaluation of Matched Back Projection

The reconstruction method with the matched back projection was evaluated and compared to the reconstruction with unmatched back projection with respect to (1) the influence on the quality of the tomogram and (2) the speed of the algorithm. The reconstruction quality was measured in terms of FWHM resolution, the SNR, and the axial elongation factor.

The reconstruction performance was evaluated by measuring the rate of convergence, i.e. the error as a function of the number of iterations. Execution times in *ms* are also provided for one exemplary hardware platform and tomogram resolution in order to allow an assessment of the performance for practical applications.

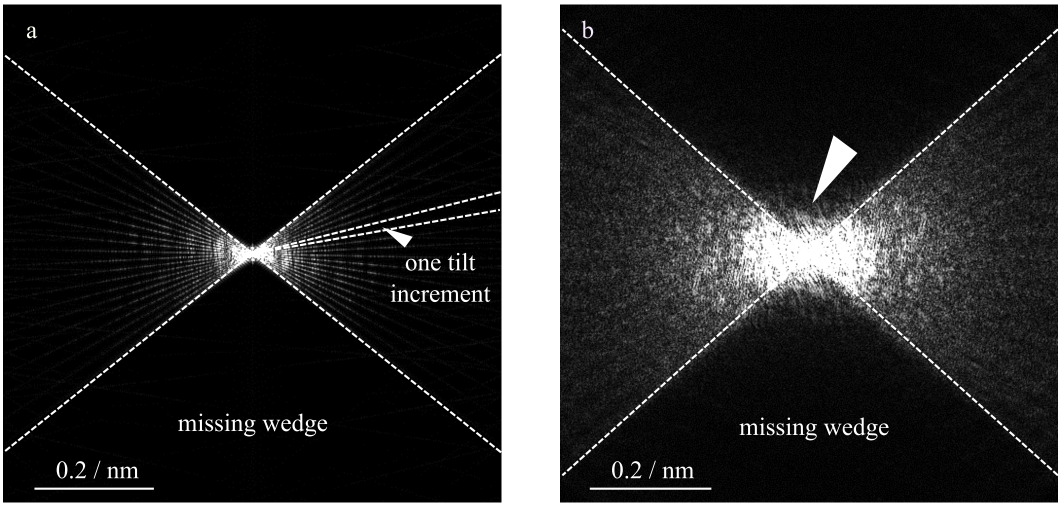

Figure 3.14: a) Frequency domain (Fourier Transform) of a xz slice of the tilt series. b) Frequency spectrum of a xz slice of the CTFS. The white lines mark the border of the missing wedge (±40°). Figure from (Dahmen et al., 2014a).

Two different error measures were used. The residual root means square error (RMSE) is defined as the root of the mean square values of per-pixel difference between virtual projection and measured projection. Residual RMSE can be measured without knowledge of a ground truth tomogram and can therefore be applied to experimental data. The ground truth RMSE is defined as the root of the mean square values of the voxel-by-voxel difference between a reconstructed tomogram and a known ground truth. This error measure is generally more expressive than residual RMSE but can obviously be measured only for reconstructions from phantom data.

The evaluation was, therefore, performed on two datasets, one experimental described above and one with phantom data. The phantom dataset consisted of eight clusters of 64 spherical particles, each of 18.4 nm diameter. The particles had a density of 16 gray scale units on a homogeneous background of density 1 gray scale unit. The clusters and the particles inside the clusters were placed at random positions. Synthetic images were generated using the combined tilt- and focal scheme and the STEM forward projection model without the simulation of imaging errors such as misalignment or aberrations. Gaussian noise with a standard deviation of 25 was added to the images.

3.3.6 Results on Experimental Data

Reconstructing the experimental dataset using the matched operator, the iterative reconstruction algorithm reached a residual RMSE of ≈ 1950 after a single iteration, i.e. after the images from all directions and all focal values were processed once. After two iterations, the method reached a residual RMSE of ≈ 1394, which was better than the RMSE of ≈ 1770 reached with the unmatched back projection after the full 120 iterations (Figure 3.11). Thus, the matched back projection resulted in a speed-up factor of about 60 to reach the same RMSE level. The reconstruction with matched back projection reached a minimum RMSE of ≈ 1214 after 20 iterations, after which it showed semi-convergent behaviour (Elfving, Hansen, & Nikazad, 2014), i.e. the error started to grow again.

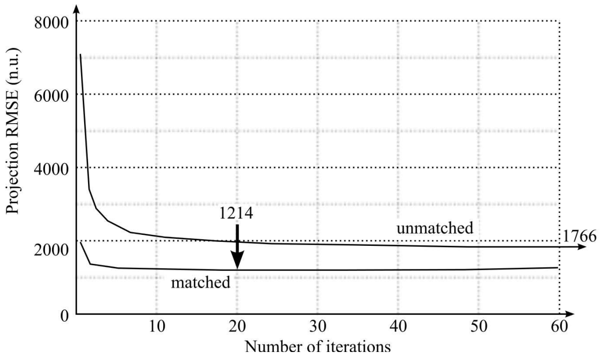

Figure 3.15: Convergence rate of the matched and unmatched back projections on experimental data. The matched operator reaches an optimum value of ≈ 1214 after 20 iterations. The unmatched operator reaches a value of ≈ 1766 after 120 iterations but the error is still marginally descending at that point. Figure from (Dahmen, Kohr, et al., 2015).

For the reconstruction with the unmatched back projection, the lateral FWHM was 9.4 ± 2.4 *nm*, and the axial elongation factor was 2.2 ± 0.5. For the reconstruction with the matched back projection, the lateral FWHM was 9.9 ± 2.5 *nm* and the axial elongation factor was 2.3 ± 0.4. Thus, the reconstructions with the matched back projection showed marginally lower FWHM resolution.

In addition to measuring FWHM, the intensity profile of a lateral cut through the center of a nanoparticle was considered (Figure 3.16a). The matched back projection showed a maximum intensity in the center of the nanoparticle of 686 gray scale units (unmatched: 1270). Thus, the matched back projection exhibited ≈ 45% lower maximal intensities. On the other hand, the signal exhibited a much lower artifact level. In order to quantify the level or artifacts present in both reconstructions, a lateral profile through an empty part of the tomogram was considered (Figure 3.16b). The matched back projection showed an artifact level of 23.0 ± 12.9 gray scale units (unmatched: 72.7 ± 62.1). By putting the maximum signal intensity in relation to the background noise, a SNR of 29.8 was computed (unmatched: 17.7), so the matched back projection generated tomograms with clearly improved SNR.

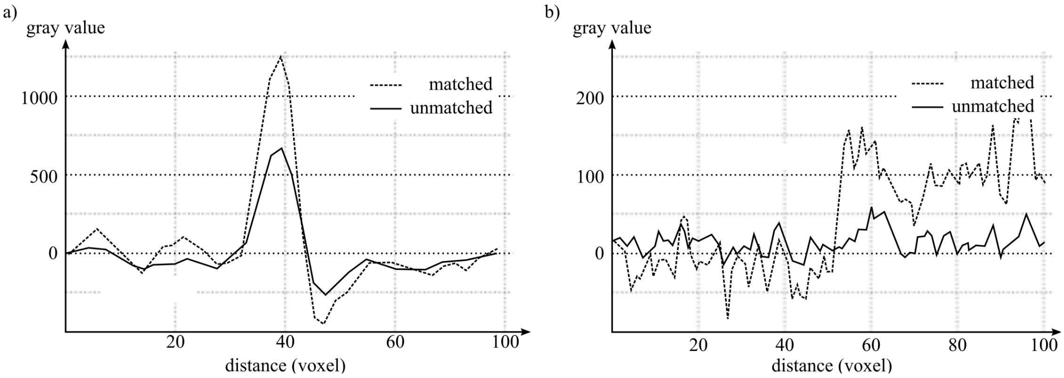

Figure 3.16: a) Profile of a lateral cut through a nanoparticle. b) Profile of a lateral cut through an empty part of the tomogram (artifacts only). Results with the matched back projection are represented by the continous line, results with the unmatched back projection are represented by the dashed line. Figure from (Dahmen, Kohr, et al., 2015).

Individual slices of the tomograms are presented in Figure 3.17 for visual inspection. As can be seen, the tomogram generated with the matched back projection looks smoother and exhibits less artifacts. The tomogram generated with the unmatched back projection looks sharper, but also noisier and with stronger artifacts.

3.3.7 Results on Phantom Data

In order to provide the rate of convergence of the ground truth RMSE, experiments were performed on phantom data. The unmatched back projection operator found a solution with ground truth RMSE of ≈ 7.3 after 30 iterations, while the matched operator reached a minimal ground truth RMSE of ≈ 7.4 after 10 iterations (Figure 3.18).

Individual slices of the reconstructions with both matched and unmatched back projection are shown in Figure 3.19. The results support the conclusion that reconstructions with the matched back projection result in smoother

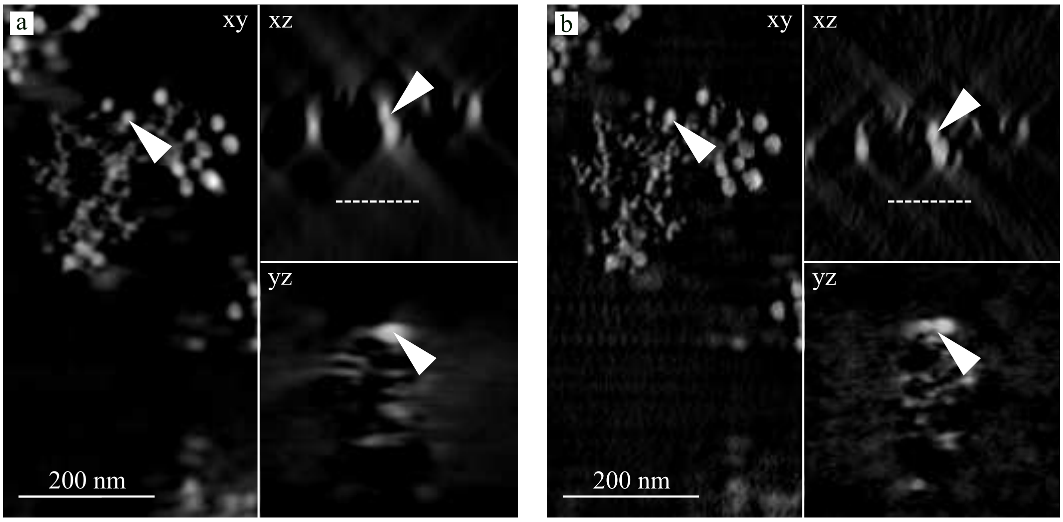

Figure 3.17: Individual slices of the reconstructions from experimental data. a) Reconstruction with the matched back projection. b) Reconstruction with the unmatched back projection. The matched reconstruction seems to give smoother results, while the unmatched back projection seems to result in sharper, but noisier images with more artifacts. The dashed lines show the position of the profiles used to determine background artifact intensity (Figure 3.16b). Figure from (Dahmen, Kohr, et al., 2015).

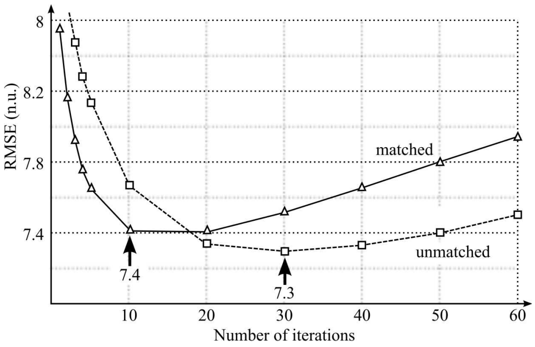

Figure 3.18: Convergence experiment on phantom data. The convergence was measured as RMSE relative to the ground truth. The arrows indicate minimal values. Figure from (Dahmen, Kohr, et al., 2015).

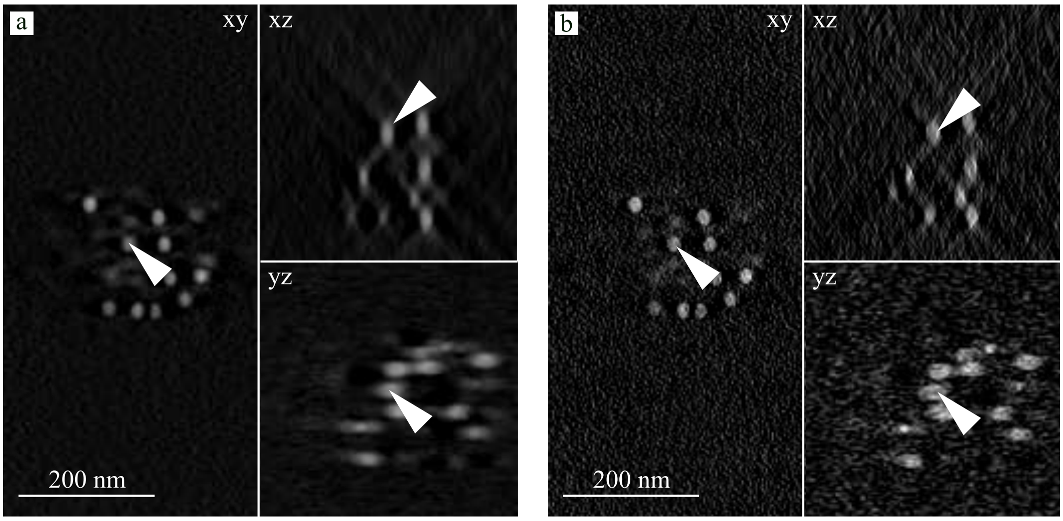

Figure 3.19: Individual slices of the reconstructions from phantom data. a) Reconstruction with the matched back projection b) Reconstruction with the unmatched back projection. Figure from (Dahmen, Kohr, et al., 2015). The reconstruction with the matched back projection results in smoother solutions with less artifacts. The reconstruction with the unmatched back projection results in slightly sharper looking solutions, but also noisier and with more artifacts.

tomograms with less artifacts. As observed on experimental data, reconstructions with the unmatched back projection look sharper but also noisier and with more artifacts.

3.4 Discussion of the Reconstruction Algorithm

3.4.1 Computational Cost per Iteration

In Section 3.3.6, a performance comparison of matched and unmatched back projection was given in terms of the rate of convergence, i.e. the remaining error or residual norms was investigated as a function of the number of iterations. This raises the question of the computational cost of a single interation. However, the total cost of the reconstruction algorithm is strongly dominated by the cost of the forward projection, which takes about 8 – 20 times longer than the back projection (≈ 960 ms on a Tesla C2075). It is therefore justified to discuss reconstruction performance mainly in terms of number of iterations.

In the following, we will nevertheless discuss the difference in computational cost for the execution of the back projection itself for completeness.

The matched back projection is considerably more costly to evaluate in comparison to the unmatched back projection. This is due to the fact that (1) the matched back projection has to compute additional convolution operations for the pre-filtering of residuals and (2) the unmatched back projection does not change every voxel in a single iteration and therefore requires access to less memory. Concrete execution times depend on the used hardware and the size of the reconstruction volume, in particular whether it fits entirely into the GPU memory or has to be processed sequentially in chunks. On a Nvidia Tesla C2075, a single execution of the matched back projection to a volume resolution of 512^3 voxels took ≈ 120 *ms*, while an execution of the unmatched back projection took ≈ 50 *ms*.

The implementations of both matched and unmatched back projection were not fully optimized. The matched back projection could be optimized by using a faster implementation of the convolution operation switching between implementations in real space and in Fourier space depending on the size of the PSF and a hardware-specific threshold. The unmatched back projection could be optimized by implementing a spatial subdivision scheme and restricting the execution of the kernel to those parts of the volume that have non-zero correction.

3.4.2 Classification of the TF-ART Algorithm

In order to understand properties of the reconstruction algorithm TF-ART, it is of interest to classify the algorithm with respect to existing classification schemes. More specifically, if we can show that TF-ART is the instance of an algorithm class that has been investigated in related work, we could gain knowledge about all proven properties of the class by "inheritance".

However, the TF-ART algorithm is an instance of the variable block ART algorithmic scheme (Censor, 1990), generalized to feature an unmatched projection / back projection pair as discussed in (Zeng & Gullberg, 2000). It can be expressed in the form:

$$X \leftarrow X + \lambda \sum_{k=1}^{m} w_k A_{back}^T (B - A_{forward} X) \tag{3.4}$$

Hereby, X is the volume, B_k are the pixels in all of the measured projections, $A_{forward}$ is the matrix expressing the forward projector, A_{back} is the matrix expressing the back projector. In this notation, which is adapted from the original paper on Variable Block ART (Censor, 1990), w_i are weights that control at which granularity and in which order the measured projections are processed, not to be confused with the discrete approximation of the basis functions. In the case of TF-ART, w_i are chosen such that they form the update loop described in Section 3.2.4. For the special case $A_{forward} = A_{back}$, i.e. in the case of an matched projection / back projection pair, the algorithm falls in the class Variable Block ART and all existing knowledge on that class applies. This includes the proof of convergence (Wang et al., 2007; Qu, Wang, & Jiang, 2009; Yan, 2010), an understanding of the incremental convergence of spatial frequencies (Norton, 1985), and the semi-convergence behavior under noisy conditions (Elfving et al., 2014) among others.

However, the implication of the generalization to an unmatched projection/back projection pair implies that existing knowledge on the algorithm classes Variable Block ART (Censor, 1990) does not apply without further effort to TF-ART. Intuitively it seems likely that TF-ART converges to the optimum of the L2 norm as well, but a proof depends on properties of the back projection as discussed in (Zeng & Gullberg, 2000).

3.4.3 Influence of the Matched Back Projection

The adjoint back projection resulted in a drastically increased rate of convergence. On experimental data, the convergence rate was about 60 times faster then with the unmatched back projection. The choice of the back projection had an impact on the solution, i.e. a different solution was found depending on which back projection was used. The solution found by the unmatched back projection looked sharper but also noisier and with more artifacts. The solution from the adjoint back projection looked smoother but less sharp. When measuring resolution and axial elongation factor using the FWHM metric, both solutions reached the same values within measurement precision, so a preference for one solution or the other is likely to depend on a given application.

3.4.4 Conclusions of Experimental Evaluation

In conclusion, tomograms generated with the CTFS recording scheme show clear and relevant improvement in the axial elongation factor, compared to tomograms generated from a pure tilt series. Thus, we demonstrate that recording scheme improves on one of the two most important types of artifacts that currently limit STEM tomography.

Two different back projection operators were investigated. Compared to the unmatched back projection, the main advantage of the matched back projection is a drastically improved convergence rate. On experimental data, the convergence rate is about 60 times higher compared to the unmatched back projection, which brings the reconstruction times to regions that are feasible for practical application. Also, the reconstruction from the matched back projection is overall smoother and exhibits a better SNR.

Chapter 4

Software Architecture

4.1 Existing Software Packages

A large number of both open source and commercial software packages for tomographic reconstruction in electron microscopy already exists. In terms of the research of this thesis, this raised the question if the proposed methods should be implemented as extension to an existing software package or if an own, new package should be implemented.

In the following, an overview of existing open source and commerical software packages for electron tomography is given. Next, key requirements are identified and matched against the existing implementations. It is argued that all existing packages fail to fullfill some of the requirements from the architecture point of view and a novel software package, called "Ettention" is presented.

4.1.1 Open Source Packages

The package IMOD (Kremer et al., 1996) is a comprehensive and established package for electron tomography processing and reconstruction, providing the weighted back projection algorithm and SIRT. IMOD provides mature CPU and cluster parallelization options, offering a framework for running generic parallel processes. Support for GPU processing using CUDA has been added more recently. IMOD is designed as a set of command-line utilities that are optionally connected through a Java-based user interface into pre-defined workflows. Even though it is possible to extend IMOD with additional reconstruction techniques, re-use of the existing code is not straightforward. This is especially true for the GPU-based reconstruction algorithms, which are designed as monolithic CUDA kernels.

The ASTRA toolbox (Aarle, van der Maar, Batenburg, & Sijbers, 2008) uses the MATLAB environment to provide GPU based algorithms for 2D and 3D tomography. It supports a fan and parallel beam geometry for 2D, and a parallel and cone beam geometry for 3D. GPU implementations are available for SIRT, Conjugate Gradient for Least Square and Feldkamp-Davis-Kress algorithm, among others. A number of methods incorporating prior knowledge are also implemented, for example the FISTA algorithm (Beck & Teboulle, 2009) for total variation minimization and the discrete algebraic reconstruction technique (DART) algorithm (Batenburg & Sijbers, 2007) for discrete tomography. Similarly to Ettention, ASTRA provides direct access to GPU-accelerated forward projection and back projection operations allowing their easy integration into algorithms implemented in MATLAB. The main shortcoming of the ASTRA toolbox from the point of view of this thesis is the coupling with the MATLAB environment, which severely limits the usability as a generic library and the possibility to interface with 3rd party software.

The software package TomoJ (Messaoudii, Boudier, Sanchez Sorzano, & Marco, 2007) is implemented as a plug-in to ImageJ, a widespread image processing tool implemented in Java. TomoJ provides functionality for the preprocessing, alignment, and tomographic reconstruction of electron tomography tilt series. The main advantage of the software is the integration in ImageJ with its powerful and well-structured application programming interface (API). Disadvantages are low performance and memory management limitations because of the use of the Java virtual machine as well as the lack of GPU support.

A strongly integrated approach to electron tomography is taken by the software EM3D (Ress, Harlow, Marshall, & McMahan, 2004). The software provides standard support for alignment and reconstruction of electron tomography tilt series. Additionally, it features advanced functions for segmentation, iso-surface extraction, and rendering of tomograms.

The Tomo3D package (Agulleiro, Garzon, Garcia, & Fernandez, 2010) takes a different approach of accelerating iterative reconstruction techniques. They exploit the parallelism on x86 architectures by distribting the executing to different cores and simultaneously making use of single instruction multiple data (SIMD) parallelism on the instruction level. The packages implements weighted back projection and SIRT and claims to reach performance comparable to highly optimized GPU implementations.

PROTOMO (Agulleiro & Fernandez, 2011; Winkler, 2007) is a software package design specifically for electron tomography. It focuses on a marker-free alignment and variable-weight weighted-back projection reconstruction. Similar to most other packages in this area, it lacks any GPU support and is built as a monolithic software.

The software package OpenMBIR introduces a unique iterative reconstruction algorithm with a forward model that includes a compensation for Bragg scattering (Venkatakrishnan et al., 2013).

UCSF Tomography (Zheng et al., 2007) was originally a software package for tilt-series collection, which made use of sample movement prediction avoiding the need for additional focusing and tracking. Recently it has been extended with real-time alignment and reconstruction with the goal of providing an immediate feedback during the data collection rather than highest-possible quality reconstruction. The real-time automatic operation is achieved by employing marker-free alignment and sequential cluster-based weighted back projection reconstruction.

Several package with focus on averaging techniques like sub-tomogram averaging and single-particle averaging also provide tomogram reconstruction capabilities as by-features. Examples include EMAN2 (Ludtke, Baldwin, & Chiu, 1999; Tang et al., 2007), SPIDER (Shaikh et al., 2008), RELION (Scheres, 2012), Bsoft (Heymann, Cardone, Winkler, & Steven, 2008), and PyTom (Hrabe et al., 2012).

4.1.2 Commercial Packages

Apart from open-source and research-based software packages, there are only a few commercial tools. IMAGIC (van Heel & Keegstra, 1981) is a general purpose, interactive image analysis software package written in FORTRAN 77. It is mainly applied in the field of high resolution biological electron microscopy. Advanced techniques are available for the analysis of images of single particles, including pattern recognition. The main current 3D reconstruction algorithm in this software is the exact filter back projection algorithm (van Heel, Harauz, Orlova, Schmidt, & Schatz, 1996).

Inspect3D (Schoenmakers, Perquin, Fliervoet, Voorhout, & Schirmacher, 2005) is an integrated software package for acquisition, alignment, reconstruction, and visualization of electron tomography data. Its key feature with respect to this thesis is the highly-optimized GPU SIRT technique. In general the package has a solid GPU parallelization base. It primarily targets electron microscopy end-users without exposing any of the GPU code blocks.

4.2 Requirements: Extensibility, Modularity, and Performance

The basic iterative reconstruction methods approximately solve the linear system $Ax = b$, where x is the searched for tomogram, b is the images from the microscope and A the system matrix modeling the imaging process. The system is both ill-posed and under determined because of imaging errors such as alignment problems, aberrations and noise. As a consequence, model based approaches, a-priori information and regularization techniques are key to high quality reconstruction. By their definition, these techniques are difficult to design in a general way and have to be tuned to the specific application scenarios and conditions. We identify three key requirements for an electron tomography software package:

- **Extensibility** The software must be extensible with respect to different image acquisition schemes, microscope geometries, file formats and reconstruction algorithms. Likely extensions also include different choices of volume basis functions, regularization methods and different means to incorporate prior knowledge.
- **Modularity** The software should be modular in the sense that orthogonal features can be recombined as easily as possible. For example, once a new reconstruction algorithm is implemented, it should work with any microscope setup, any input file format, and ideally with any choice of volume basis functions.
- **Performance** The software should deliver good performance with respect to two different aspects. Obviously, reconstruction times should be low for the convenience of the user. Equally important however is he scalability to high resolution data, i.e. the software must be capable to process high projection and volume resolutions within an acceptable time frame.

By comparing those requirements to the existing packages, one easily sees that most packages excel at an individual requirement, but no single package matches all three simultaneously. For example, the IMOD software delivers highly optimized performance for high resolution volumes, but does not feature the required level of modularity. The ASTRA toolbox, on the other hand, is highly extensible and modular, but restricted to low resolution data.

Therefore, a new software package, called "Ettention" is proposed. Ettention consists of a set of building blocks than can be adapted and combined to application specific iterative reconstruction methods. In this respect, the ideas behind Ettention are similar to the concepts driving the design of the ASTRA toolbox with the difference, that Ettention additionally aims for immediate usability by microscopy end-users by providing GPU support for high resolution data.

4.3 The Ettention Architecture

4.3.1 Overview

In order to support a large variety of different high performance computing (HPC) hardware architectures, the OpenCL API (Khronos Group, 2007) was used as an abstraction layer. While this solves the primary problem (i.e. the system can run on different hardware architectures) the OpenCL programming model still exposes a number of technical properties, such as the need to very explicitly handle parallelism and memory management. Those needs are addressed by a hierarchy of abstraction levels as explained in the following (Figure 4.1).

At the lowest, most technical level are primitive operations such as volume ray tracing operations. The next layer are kernel classes that contain an individual OpenCL Kernel and a C++ class that wraps the kernel and serves as an interface to the kernel. A call to the "execute" method of the kernel class corresponds 1:1 to an execution of the kernel on the HPC device.

However, because the reconstruction volume does not necessarily fit into HPC device memory, many operations include data transfer of parts of the volume, processing by consecutive kernel executions and accumulation of results. These aspects are handled on the next abstraction layer, the operator level. Operators are self-contained concerning memory management and parallelism. Simple operators contain exactly one kernel class and use consecutive executions of this kernel to perform their function. The most important simple operators are the generic projection and back projection operators. They can be combined with different forward and back projection kernels to support a variety of projection geometries and basis functions. Complex operators recursively combine other operators to implement functionality. The most important complex operators are reconstruction algorithms such as a generic block iterative algorithm and the special cases SART operator and SIRT operator.

Apart from the performance critical parts that are implemented on the HPC device, Ettention is implemented in C++ and runs on the CPU. The C++ part also contains a plug-in interface, such that the software can be

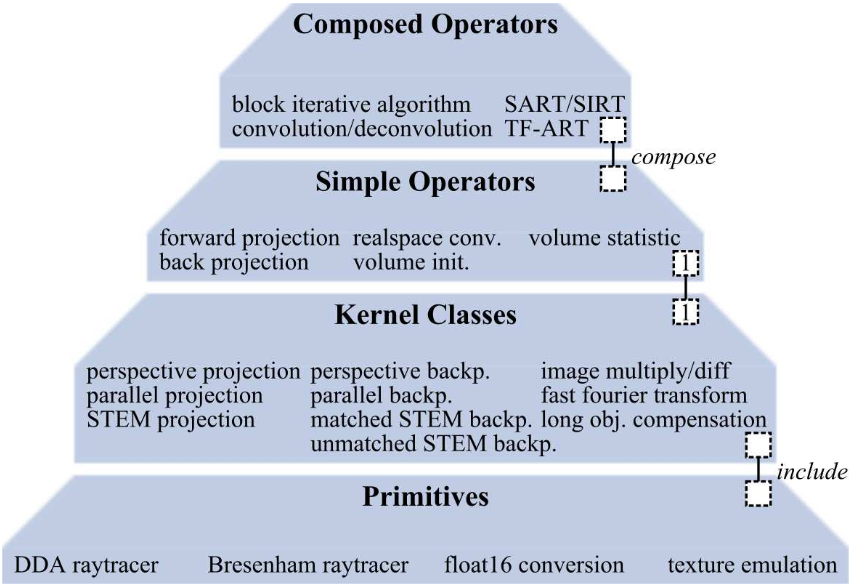

Figure 4.1: The hierarchical view of the HPC part of the Ettention architecture consists of four layers: primitive operations such as volume ray tracing are implemented as OpenCL includes. Kernel classes wrap exactly one OpenCL and serve as interface to this kernel. Simple operators are are self-contained concerning memory management as they perform data transfer to HPC memory as required. Composed operators hierarchically combine other operators.

extended in almost every aspect by writing plug-ins. The interface for plug-ins allows to instantiate the following concepts:

- reconstruction algorithms
- forward projections
- back projections
- file formats (input and output)
- data access schemes
- microscope and projection geometries

In summary, Ettention is implemented as a modular set of HPC building blocks. The individual blocks form a hierarchy of abstraction, such that the technical aspects of memory management and parallelism are hidden in the lower layers. At the higher level of abstraction, objects called “operators” can be used without regard of the HPC aspect. The system can be extended via plug-ins in almost every part, and the combination of extension often works out of the box.

4.3.2 Plug-in Concept and Binaries from Kernel Files

Ettention is structured in a central library that can be extended by plug-ins that are loaded at runtime via dynamic linking. This helps for separating the source code into manageable units and allows the framework to be used by developers from organizations with different requirements regarding their licensing terms. Particularly, it offers the possibility to provide closed-source, commercial plug-ins to the open source Ettention framework. In order to build self-contained binary files and allow closed source developers a minimal level of protection on their compute kernel code, Ettention implements a mechanism to link compute kernels into the executable binary. As the kernel source code is automatically embedded as a string constant at build time, obfuscation could trivially be integrated into the build process, if required.

4.3.3 Memory Management

A typical reconstruction volume has a size between 512^3 and $4096^2 \times 1024$ voxels, at typically 16 or 32 bit floating point precision for gray values. This implies a memory requirement of 512 mega bytes (MB) to several giga bytes (GB) for the volume alone so it cannot be assumed that the volume fits into the limited HPC memory, and out-of-core mechanisms for GPU need to be applied in those cases. As a consequence, a reconstruction operation such as a forward projection does not map 1:1 to the execution of a HPC compute kernel.

One approach is to reduce the 3D reconstruction problem to a series of 2D reconstructions, i.e. slice both the volume and the projection along a plane perpendicular to the tilt axis and reconstruct a slice at a time. While providing clear advantages in terms of simplicity and reconstruction performance, the method has the obvious disadvantage of restricting projection directions. For Ettention, we decided not to take this approach and allow arbitrary projection directions and geometries. This enables use-cases such as laminography (Maisl, Porsch, & Schorr, 2010), STEM tomography with convergent beams (Dahmen et al., 2014a), cone-beam tomography (Mueller, Yagel, & Wheller, 1999), or in the future tomography from unaligned stacks with arbitrary projection matrices.

As a consequence, most operations involve sequential processing of the volume including consecutive upload of parts of the volume to the GPU, execution of a kernel that operates on a subvolume, and accumulation of results. Those operations are performed transparently inside the individual building blocks, so the memory management aspect of GPU programming is hidden when developing using Ettention. The handling of GPU memory management is performed using a template called "GPUMapped<>", which represents the notion of an object (such as a reconstruction volume, image, or projection matrix) that is required both on the CPU and the GPU. The template links the CPU object to a configurable GPU representation, such as a float buffer (16 or 32 bit precision) or a GPU image (1D, 2D, or 3D layout; 16 or 32 bit precision). By tracking the up-to-date status of the different representations, the system can figure out when data transfer to device memory or back is required.

4.3.4 Parallelism

One core concept when programming a system that heavily relies on the GPU is the need to explicitly handle parallelism and synchronization of concurrent results on HPC devices. While reconstruction algorithms exhibit a high degree of parallelism at the lower levels, the high-level formulation of the algorithms are basically sequential. Ettention addresses this situation by introducing algorithmic building blocks that operate on problem domain objects, such as volumes and projection images. The building blocks are implemented in a parallel way and are self-contained concerning memory management and synchronization in the sense that they keep track of memory status and perform data transfer to and from GPU as required. They can be recombined to formulate reconstruction algorithms in a sequential way using operations such as image loading, forward projection, or back projection.

However, using predefined building blocks is not sufficient for most innovative applications. Introducing novel building blocks for new compute tasks will require dealing with platform specific aspects at some point. However, most building blocks required for reconstruction algorithms fall in either of the three categories forward projection, back projection, or image manipulation. They differ mainly in the innermost loop, for example by a custom regularization method, new volume basis functions, or custom projection geometry. Because those features are implemented one abstraction level below the per-pixel or per-voxel level that exposes the parallelism, the concepts can easily be expressed in a way that is agnostic to parallelism.

However, having function calls in the innermost loops would be prohibitively expensive and therefore not exposed by HPC devices at all. Instead, Ettention solves this by injecting statements in the OpenCL source code via macros. This way, custom regularization functions or new basis functions are inlined and optimized by the kernel compiler and therefore do not impact performance.

4.4 Individual Building Blocks

In the following, the most important algorithmic building blocks are presented. We follow a bottom up approach starting with the primitives and moving to concepts of successively higher levels of abstraction.

Grid Traversal

A ray traversal for uniform grids is provided as a fundamental operation to most forward projection operations. The ray traversal is based on a 3D version of the digital differential analyzer (DDA) (Watt, Allan, 2000) algorithm but uses floating point operations instead of the fixed point arithmetic of the original implementation for better performance on HPC devices. This module is mainly used to implement forward projections.

A separate version of the ray traversal exists that introduces the notion of a ray thickness, such that it iterates over all voxels within a given distance to the ray. The version is implemented using a modified 3D Bresenham (Bresenham, 1965) algorithm. It is used for tracing rays through volumes assuming basis functions that span more than one voxel, i.e. blobs (Marabini et al., 1998; Bilbao-Castro et al., 2009).

Forward Projection

Forward projection operations are provided as self-contained building blocks for reconstruction algorithms. Versions are provided for parallel projection such as applicable to TEM tomography and for perspective projections, usable for X-Ray computed tomography (CT). A version assuming parallel projection but limited depth of field is provided for use with aberration corrected STEM. The forward projection is implemented pixel-by-pixel and based on the grid traversal routines. The forward projections are provided in variations for voxel and blob basis functions, using the ray traversal routines based on DDA or Bresenham as explained above.

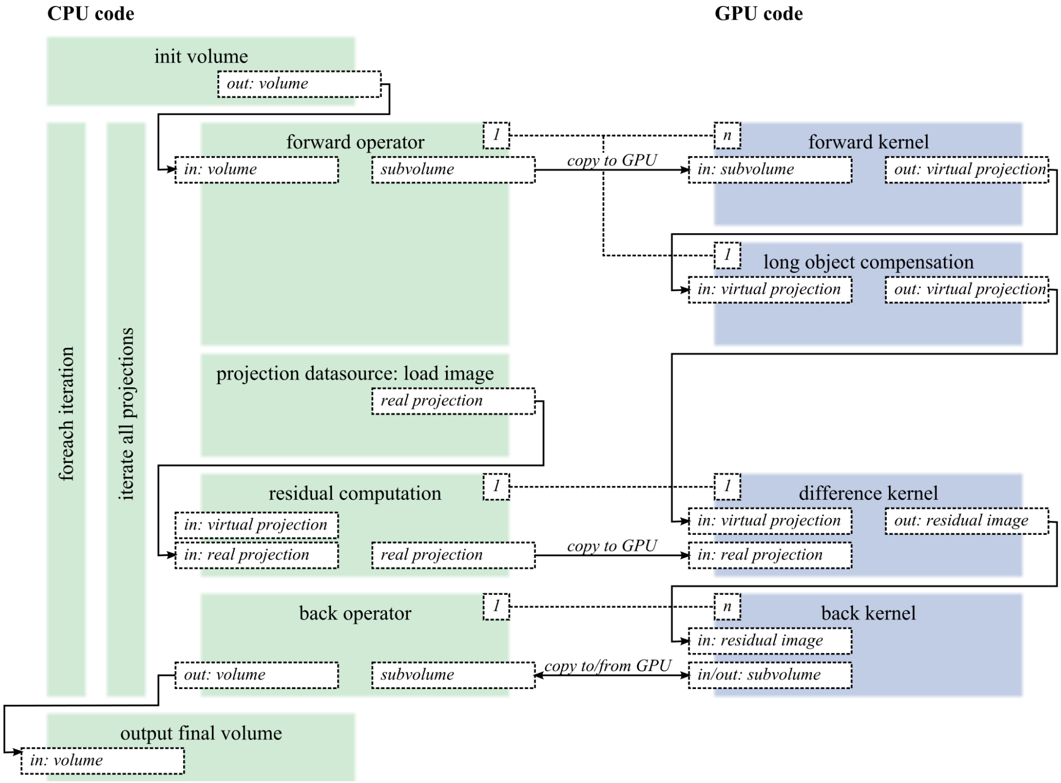

Figure 4.2: A block diagram of the SART algorithm as an example for an iterative reconstruction algorithm implemented using the building blocks in Ettention. CPU code objects including data flow are shown on the left side (green), GPU kernels on the right side (blue). Some operators like the residual computation correspond 1:1 with calls to GPU kernels, while the forward and back projection operators, that work on volumes, require more than one call to the corresponding kernel. A hierarchical mapping, where one operator uses several different kernels is also possible and shown on the example of the forward operator, which uses several calls to the forward kernel and an successive, optional call to the long object compensation. Most input or output slots of the building blocks are implemented using the GPUMapped<> template, so the operator can either take CPU data input, or it can take a buffer that already lies in GPU memory space. An example for this is the virtual projection input for the difference kernel, which already is on the GPU, so the system transparently figures out that a copy is not required. Figure from (Dahmen, Marsalek, et al., 2015).

Back Projection

Building blocks for back projection operation correct a volume by projecting each voxel back onto an image plane and adjusting the voxel intensity using a customizable regularization function applied to the residual. The operation is implemented in a voxel-by-voxel way in order to avoid no synchronization issues for writing updates to the volume. This results in much higher performance than a pixel-by-pixel implementation using ray traversal. A discussion of the topic can be found in (F. Xu & Mueller, 2005).

Because the projection geometry is provided as a general 4x4 matrix, the operation can be used with a large variety of microscope geometries. Preconfigured building blocks are provided for unregularized parallel and perspective back projection and several regularization schemes, for example for use in STEM tomography (Dahmen et al., 2014a).

Image Processing Operations for Residuals and 2D Projections

Several building blocks are provided that operate on individual images. Because Ettention assumes several individual projections to fit into GPU memory, the per-image operations can directly be mapped to one execution of an OpenCL kernel. Per image operations include the computation of residuals (implemented as a per-pixel difference computation) and long object compensation (W. Xu et al., 2010), real and complex multiplication, and several statistics functions such as minimum, maximum, and mean pixel intensity, and mean root square computation.

Data Input/Output of Multi-Dimensional Data

Ettention provides support to load the most common data formats for image stacks, such as the mrc format, multi-slice tiff, or an image sequence in a directory. The term image stack hereby refers to a 1D array of 2D images.

In some use cases, a 1D array of images is insufficient. For example in the case of a CTFS, the input consists of one image per tilt direction and focal positions, resulting in a 2D array of images. Ettention therefore introduces the concept of a “hyperstack”, that consists of an n-dimensional array of 2D images. Images in a hyperstack are referenced using n-tupel of integers as indices.

The reconstruction algorithms generally work on such hyperstacks. This way, the only part of the algorithm that depends of the dimensionality of the input is the projection ordering, called ProjectionSetIterator in Ettention terminology. For example, the multi-level access scheme (Guan & Gordon, 1994) for single-tilt tomography assumes a 1D stack. But given a suitable way to iterate, a reconstruction algorithm can operate on input of any number of dimensions.

Generally, in Ettention, it is assumed that individual images fit into memory, but not necessarily the entire stack. As a consequence, iterative techniques might require loading input images several times from disk. In Ettention, the loading of images from files is therefore implemented via an interface called "datasource" that allows random access to individual images (instead of a deserializer or stream operator). An in-memory representation of a hyperstack is missing intentionally, instead the API provides a caching datasource, which will keep a configurable number of images in memory using a least-recently-used strategy, and load missing parts of the stack as required.

Fast Fourier Transform, Convolution, Deconvolution

A very useful basic operation is the transfer of images from real space to frequency space and back by means of a Fast Fourier transform (FFT). On the OpenCL platform, the clAmdFFT library provides a highly optimized FFT implementation. However, due to its very flexible and technical interface, integrating clAmdFFT into a given application is tedious. Ettention encapsulating the library as a read-to-use building block that operates directly on GPUMapped<> images.

One application of the FFT is the intermediate output of buffers for debugging purposes. Any buffer can be configured to be written to disc at configurable points in the algorithm (per iteration, per projection, or per kernel execution). The output can be performed either in real space or in frequency space, which is useful during algorithm design and debugging.

Another application of the FFT is the implementation of convolution and deconvolution operations. For filter kernels with a diameter above a hardware specific threshold (20 – 60 pixels on most GPU platforms), those operations are most efficiently implemented as FFT, followed by a complex multiplication (convolution) or multiplication by the inverse (deconvolution), and a second FFT to come back to real space. A deconvolution is required as component of many reconstruction algorithms, such as weighted back projection and TF-ART (Dahmen et al., 2014a).

Reconstruction Algorithm Abstraction

With the building blocks described above, one can now easily implement novel reconstruction algorithms. Because most of the technical issues of HPC programming are hidden by the reconstruction building blocks, the algorithms themselves can be implemented in a language that clearly reveals the algorithmic idea. The reconstruction algorithms work on interfaces such as generic forward projection or generic back projection. Therefore, the individual building blocks can be exchanged. This allows a large variety of combinations and is key to the requirement "Modularity" that was claimed in the introduction.

Ettention provides a generic block iterative reconstruction operator (Censor, 1990). This operator allows to configure the block size and works with arbitrary forward- and back projection operators, potentially provided via plug-ins. In this framework, the well-known algorithms SART and SIRT can be realized by initializing the block iterative operator with a parallel forward and unregularized back projection and a block size of one (SART) or equal to the number of projections (SIRT). Different schemes can be implemented via plug-ins.

4.5 Interfacing with Other Software

Ettention was designed as a library that can be linked into various applications. It also provides a command line interface for manual usage. Therefore, two basic scenarios exist to interface Ettention with other software. One can either link the library as a dependency into a system like a graphical user interface program or web server. Required steps for adaptations such as passing of parameters or parsing of input file formats unknown to ettention can be implemented in the application in this case.

Alternatively, one can use the command line tool to integrate Ettention into a larger system and implement the required glue code as a plug-in to Ettention. The later approach can be used for the integration into programming environments that do not easily allow linking with C++ libraries, such as the Java environment.

One example for such an integration is the eTomo package by IMOD. eTomo is basically a front end written in Java providing easy access to the separate algorithmic steps, that are each implemented as individual executables. The algorithms are called with a set of parameters stored in config files, with "tilt.com" being the basic one used for all operations.

The adapter to incorporate Ettention into IMOD consists of two components. An abstract window toolkit panel provides a user interface to set all necessary parameters. It is provided as a jar file for Java and integrated into the eTomo front end. When the reconstruction action is selected, it generates a parameter file "ettention.com" and calls the Ettention command line tool. Additionally, an Ettention plug-in provides a parser for the handling of the .com files. Upon execution, the program parses two files, "tilt.com" and "ettention.com", reads the input files, performs the reconstruction and generates the output tomogram in the IMOD compatible mrc format.

4.6 Results

4.6.1 Evaluation Methods

The Ettention software package was evaluated experimentally on a dataset of human anti-HIV-1 gp120 antibody IgG1-b12. The motivation for choosing this particular dataset was that for reasons of comparability between implementations, a public available dataset from the Electron Microscopy Pilot Image Archive (Patwardhan et al., 2012, 2014) should be selected. At the time of publication, EMPIAR-10009 was the only dataset of the database that showed a structure in the native environment, as opposed to isolated particles.

The input dataset (EMPIAR-10009) consists of a single axis tilt series of 51 images, each of which has a resolution of 2048^2 pixel. The tilt series was recorded with parallel electron tomography using a saxon scheme over the tilt range of ±60° in 3° increments. Additional information on the sample preparation and image acquisition is given elsewhere (Diebolder et al., 2014).

Tomograms were reconstructed using the IMOD implementations of back projection and SIRT and the Ettention implementations of SART as well as SIRT. Tomographic reconstructions were performed with two different output resolutions to investigate the in-core case and the out-of-core case separately. A tomogram resolution of $2048^2 \times 1024$ was selected for the out-of-core case and $1024^2 \times 512$ for the in-core case.

4.6.2 Reconstruction Quality

As a baseline, the EMPIAR-10009 dataset was reconstructed using the IMOD package (Figure 4.3). Both back projection and 10 iterations of SIRT recon-

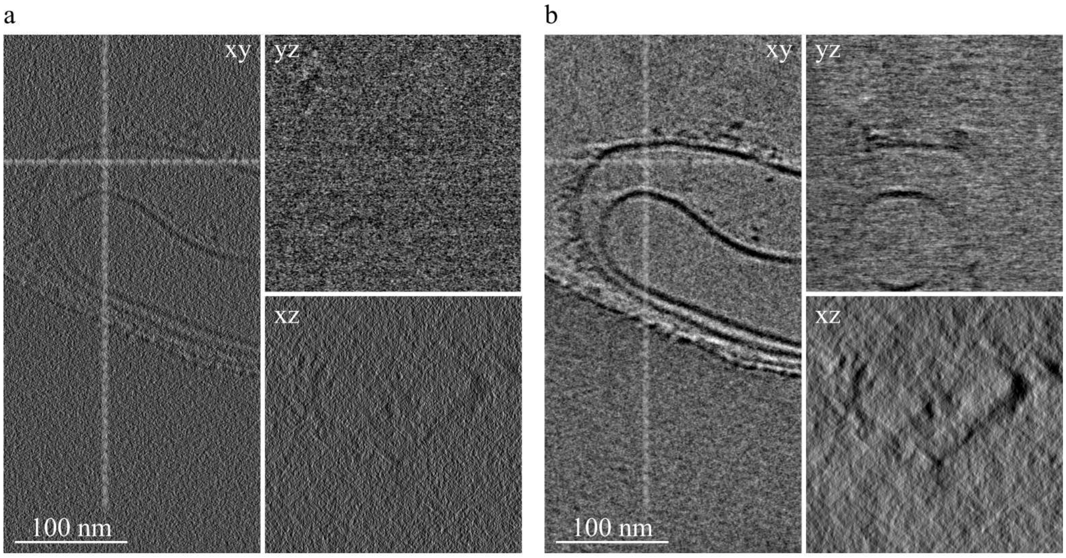

Figure 4.3: Tomographic reconstructions performed using IMOD. a) Back projection reconstruction. b) SIRT 10 iterations. The dashed lines in the xy plane indicate the relative positions of the yz and xz plane. Contrast was scaled for optimal display. Figure from (Dahmen et al., 2014a).

structions were performed. As was expected for a dataset with low SNR and limited tilt range, the SIRT reconstruction showed better contrast and revealed more details (Figure 4.3).

The same dataset was reconstructed using the Ettention software package. The reconstruction was performed using 10 iterations of SIRT (Figure 4.4a) and 3 iterations of SART (Figure 4.4b). As can be seen, the SART reconstruction reveals more details than the weighted back projection (WBP) reconstruction performed using IMOD, but looks noisier than the SIRT reconstructions performed with IMOD or Ettention. Compared to the IMOD SIRT, the SIRT reconstructions with Ettention seem to yield slightly smoother results that reveal more details, particularly in the yz plane. However, the Ettention reconstruction contains a low-frequency, dark region that is not present in the IMOD reconstruction. So Ettention generates a notably different reconstruction result than IMOD, even if both software systems use the SIRT algorithm. Reasons for this might include (1) the higher precision for the volume representation in Ettention, and (2) the fact that IMOD uses an additional weighting filter (Wolf, Lubk, & Lichte, 2014) that is not part of the original SIRT implementation. A quantitative comparison depends on the figure of merit, and therefore on the intended use of the tomogram, i.e. if the data should be used for segmentation, automated analysis, or single particle averaging methods.

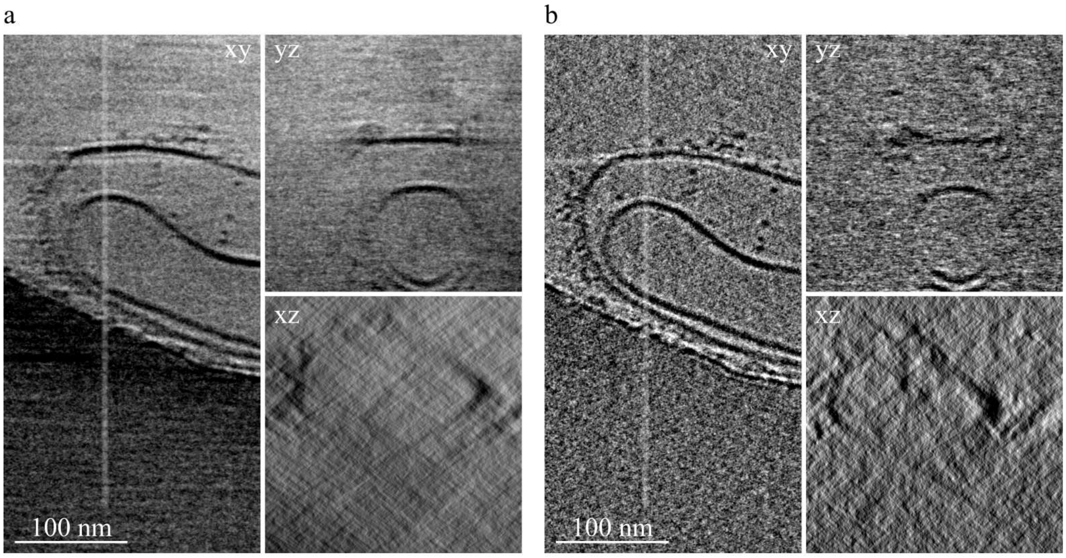

Figure 4.4: Tomographic reconstructions performed using Ettention. a) Reconstruction using 10 iterations of SIRT. b) Reconstruction using 3 iterations of SART. The dashed lines in the xy plane indicate the relative positions of the yz and xz plane. Contrast was scaled for optimal display. Figure from (Dahmen et al., 2014a).

4.6.3 Performance

Test Hardware

The performance measurements were performed on a number of different hardware platforms. A Nvidia Tesla C2075 with 6 GB memory represented a typical GPU workstation setup. A two CPU System with Intel Xeon X5560 and 48 GB memory represented a high-end CPU setup and an Intel Xeon Phi 31S1P with 8 GB memory was included to represent a non-GPU HPC platform.

Total Reconstruction Times

All reconstructions were tested on the EMPIAR-10009 dataset. The reconstructions were executed twice, once from a projection resolution of 2048^2 pixels to an output resolution of $2048^2 \times 1024$ voxel (2K case) and once from a reduced projection resolution of 1024^2 pixels to a tomogram resolution of $1024^2 \times 512$ (1K case). All execution times are given per iteration in seconds (Table 4.1).

Table 4.1: Reconstruction times per iteration in seconds on different hardware platforms.

	SIRT		SART	
	1K	**2K**	**1K**	**2K**
Tesla C2075	21 s	159 s	26 s	630 s
2 Xeon X5560	415 s	-	376 s	-
Xeon Phi 31S1P	171 s	-	246 s	-

As expected, the GPU platform shows reasonable performance, particularly in the in-core (1K) case while the CPU platform is outperformed due to inferior floating point compute power. The Xeon Phi performs better than the Xeon but does not reach the performance of the Tesla card. In the 2K (out-of-core) case, the SIRT implementation outperformed SART because the volume has to be transferred via the PCI bus less often as a result of the larger block size. However, when comparing the execution times of SART and SIRT, one has to keep in mind that the SART algorithm typically shows a higher rate of convergence, so longer runtime per iteration can potentially be compensated by using less iterations. Execution times for the 2K case on the Xeon and Xeon Phi architecture were too high to be measured in reasonable times.

Performance Analysis on Nvidia Tesla Platform

The performance on the Nvidia Tesla platform was further analyzed in more detail (Table 4.2). Forward- and back projection operator were measured separately, while the long object compensation and residual computation kernels were summed as “other kernels” as their runtime is marginal. In the 1K case, every kernel was executed 51 times, once per projection. In the 2K case, the volume was processed in 8 chunks, consequently each kernel had to be executed 408 times.

All data transfer times via the peripheral component interconnect (PCI) bus were measured separately for both directions (host-to-device and device-to-host). The volume was represented using a 3D GPU texture for the forward projection and using a float buffer for the back projection, because Nvidia platforms do not support write operations to 3D texture (missing support for cl_khr_3d_image_writes). As a consequence, the volume had to be recoded on the GPU when switching from back projection to forward projection, which is listed separately as well. The last data row “Non-HPC processing” lists

Table 4.2: Detailed analysis of the execution time on the NVidia Tesla hardware.

	SIRT		**SART**	
	1K	**2K**	**1K**	**2K**
Forward projection	4.7 s	36.47 s	4.7 s	36.5 s
Back projection	12.3 s	97.82 s	12.3 s	97.6 s
Other kernel	0.02 s	0.073 s	0.02 s	0.07 s
Data transfer to GPU	1.0 s	6.48 s	0.74 s	224 s
Data transfer from GPU	0.31 s	2.5 s	0.52 s	212 s
Encoding as GPU texture	0.12 s	0.94 s	6.0 s	48.0 s
Non-HPC processing	2.5 s	14.7 s	1.8 s	10.1 s
Total iteration time	21 s	159 s	26 s	630 s

the difference between total iteration time and the sum of all OpenCL time measurements and collectively accounts for the time spent in C++ code on the CPU including system calls for disc input/output operations, host memory management, and so on.

As can be seen, in the 1K case SIRT and SART show comparable and reasonable performance. The difference can be explained by the additional time for encoding as GPU texture, which occurs because the SART algorithm switches between forward and back projection more often and thus has to perform additional encoding operations to store the volume into texture memory.

In the 2K case, the SIRT algorithm takes a factor of ≈ 7 longer than in the 1K case. Considering that the volume has 8 times more voxel and the projections 4 times more pixel, this is actually quite good. The performance of the SART algorithm on the other hand drops drastically. This can be explained by the fact that because the volume is processed in chunks, it has to be transferred via the PCI bus once per update block (Censor, 1990) of the algorithm. This means, the SART algorithm transfers the volume once per projection over the bus, while the SIRT algorithm performs the transfer only once per iteration. The corresponding data transfer times are clearly visible in the data ("Data transfer to GPU" and "Data transfer from GPU") and dominate the total execution time of the reconstruction.

Table 4.3: Detailed analysis of the execution time on the Intel Xeon Phi hardware.

	SIRT	**SART**
Forward projection	$28\,s$	$52\,s$
Back projection	$128\,s$	$128\,s$
Data transfer	$4.8\,s$	$3.7\,s$
Total iteration time	$171\,s$	$246\,s$

Performance Analysis on Xeon Phi Platform

A detailed performance analysis on the Xeon Phi platform (Table 4.3) showed that the run time is dominated by the cost for the back projection. A bottleneck analysis using Intel VTune Amplifier XE 2015 (update1) revealed a high cycles per instruction rate (Intel Corp., 2014a) of 9.4 for the back projection and a very high latency impact (Intel Corp., 2014a) of 294 units. The source for those issues could be tracked to the access of the volume representation in device memory in the correction step. As this memory access has already close to optimal layout (consecutive access in work group dimension zero), this results remains inconclusive to some degree.

The data transfer times via PCI bus were also about a factor of ≈ 2 longer than on the Nvidia platform, even though the same data representation was used (16 bit half-float). This might be an indication that there are some issues with the used driver version and operating system combination (Windows server 2012 R2 and Xeon Phi driver version 3.4.32131.0 date 8/26/2014).

4.7 Discussion of the Software Architecture

Compared to existing software packages for tomographic reconstruction in electron microscopy, Ettention fills a gap between IMOD and the ASTRA toolbox. For algorithmic research without the need for immediate application to high resolution data, using the ASTRA toolbox can be recommended because of the powerful MATLAB language binding and because the limitation to the in-core case, i.e. to small reconstruction volumes, is not relevant. For microscopy experimentalists who require well established reconstruction algorithms, using IMOD is likely the best choice because of its highly-optimized reconstruction performance on high-resolution data. For projects that require both, algorithmic innovation and immediate application to high-resolution experimental data, the Ettention software package should be considered, as it delivers reasonable (though not optimal) performance even on high resolution data and exposes a rich and well-structured API that allows efficient implementation of algorithmic innovations.

The performance of block iterative reconstruction methods on HPC platforms needs to be discussed separately for two cases. In the in-core case, it can be assumed that the reconstruction volume fits entirely in device memory. In this case, reconstruction performance is limited mainly by memory bandwidth to device memory (on GPU architectures), or computational power (on CPU architectures). In the Ettention framework, memory management is handled entirely transparent for the in-core case and it is justified for the application developer to ignore those aspects during algorithm design.

In the out-of-core case, the reconstruction volume does not fit into device memory. Memory management is still handled transparently in this case, but the volume has to be transferred via the PCI bus to the device, which becomes the factor limiting performance. Per update block in the reconstruction scheme, the transfer is required two times from host to device (once for forward and back projections each) and once back from device to host (for the result of the back projections). As a consequence, iterative schemes with large blocks (like SIRT) will result in lower runtime per iteration than schemes with small blocks (like SART), as the latter require to transfer the volume more often. The algorithmic approach to optimize in this situation would be to balance convergence rate against data transfer and choose an optimal block size. At this point, the approach of ignoring hardware specific aspects entirely when developing algorithms reaches its limits.

A different solution is to technically optimize for bus bandwidth. Potential techniques here include the asynchronous transfer of volume data to and from device memory parallel to computations. Additionally, reducing the precision of gray values from 32 bit floating point to 16 bit half float is already supported by Ettention via configuration and reduces the bandwidth requirement by 50%, but results in reduced reconstruction quality in some cases.

Limiting the bus bandwidth by data compression can additionally help to maximize transfer. Compressed data can either be deflated on the HPC device for each chunk, or a suitable compression scheme can be used that allows efficient random access, i.e. supports working directly on compressed data. The later typically reaches lower compression rates but has the additional advantage that HPC caches benefit from data compression as well.

Techniques in the direction of data compression have been investigated in the field of compressed direct volume rendering (Rodriguez et al., 2013) and could be applied to tomography in those cases where the compressed data scheme still allows changing volume values. From a software architecture point of view those approaches raise the question how the proposed techniques can be incorporated in a transparent way, i.e. without complicating the code of reconstruction algorithms based on the framework. This should be addressed in future work.

As expected, performance in x86 architectures is sufficient for tests runs with low resolution but for high resolution reconstructions HPC platforms should be used. Note that this statement refers to OpenCL code and a different result might be achieved by a native implementation with platform specific optimizations, such as specialized SIMD CPU code (Agulleiro & Fernandez, 2011).

A special case is the performance measurement on the Intel Xeon Phi platform. While it was made sure that the basic performance optimization rules for the platform (Intel Corp., 2014b) were not violated, no major design changes were made to Ettention in order to optimize for the Xeon Phi. Despite close to optimal memory access patterns, the performance on this platform was restricted by the memory bandwidth to device memory. The main optimization direction in the future will therefore be a more compact volume representation. Further investigations seem justified in this regard.

As mentioned before, the main motivation for the Ettention framework is to provide researchers a development platform for the rapid prototyping of new reconstruction algorithms while at the same time allowing reasonable performance as well as integration in existing standard software at the same time. In order to facilitate this platform approach, Ettention is released under a GNU Lesser General Public License (LPGL) and can be downloaded freely from www.ettention.org. Ettention can be extended using the plug-in approach described in Section 4.3.2, and plug-ins can have arbitrary licensing models, i.e. it is allowed that third parties write closed source Ettention plug-ins under commercial licenses.

Chapter 5

Conclusions

In this thesis, tomography using STEM with limited depth of field was investigated. In the beginning of this thesis (Section 1.1), a number of research questions were stated. In the following, answers to those questions are presented as the concluding statements of the thesis.

What are the implications for the theory of tomographic reconstruction when considering images that are recorded with limited depth of field (DOF)?

The "STEM transform" is introduced as a new forward projection model that takes the limited DOF of aberration corrected STEM into account. It extends the ray model of the electron beam to a double cone, such that the convergent nature of the electron probe in aberration corrected STEM is considered. The operator is investigated analytically and it is shown that it is (1) a linear convolution, (2) a generalization of the ray transform for parallel illumination that contains the latter as the special case $\alpha \to 0$. A central contribution is the insight that (3) the STEM transform is self-adjoint.

How can the tomographic reconstruction problem be solved for data recorded with a CTFS?

Based in the theoretic model of the STEM transform, we introduce an iterative reconstruction algorithm based on Kaczmarz method. The algorithm uses a software implementation of the STEM transform as a forward projection. The implementation is based on a cone tracing implementation using stochastic ray tracing and stratified rejection sampling. Two different back projections are implemented and investigated. The first one, called "unmatched" is based in a heuristic weighting factor. The second back projection is called "matched". It uses the theoretic result that the STEM transform is self-adjoint. An efficient implementation of the adjoint operator is presented based on precomputed convolution operations and linear interpolation.

It is shown experimentally that the CTFS leads to a significant reduction of the blurring artifacts called "axial elongation" and thus to a more truthfull representation of the 3D shape of objects compared to a tilt series. The matched back projection results in a convergence rate that was ≈ 60 times higher than observed with the unmatched back projection and comparable a SART reconstruction of tilt series data.

Is there a statement comparable to the Fourier slice theorem that applies to a combined tilt- and focal series?

The Fourier transform of the STEM transform was computed analytically elsewhere (Intaraprasonk et al., 2008, Equation 27). It covers a region in Frequency space that corresponds to all but a double cone of opening angle $\pi - \alpha$, where α is the opening angle of the electron beam. This has interesting implications, as the STEM transform allows to cover the entire frequency space with a finite number of projections. Furthermore, the double cones overlap in the regions corresponding to the highest frequencies.

How should a software architecture be designed to handle data recorded with a CTFS?

We identified three key requirements for a software architecture for iterative tomographic reconstruction algorithms: Extensibility, Modularity, and Performance. The software package Ettention is presented to address these demands for a wide range of tomographic reconstruction problems, including the reconstruction of CTFS data. The software consists of building blocks, which can quickly be assembled to application specific tomographic reconstruction algorithms.

The software package can be extended in almost every aspect using plugins. The Ettention framework hides the technically challenging aspects of HPC programming by providing an abstraction layer above memory management and parallelism, enabling the formulation of reconstruction algorithms in a domain-specific language. The system provides a set of building blocks, called operators, that can freely be combined to form new reconstruction algorithms in a modular way.

Ettention provides feasible performance for a wide range of HPC platforms and allows for the integration of reconstructions into existing software solutions. We propose Ettention as a platform for algorithmic research in situations that require both rapid prototyping of algorithms and application of the new methods to experimental high-resolution data sets, respectively the integration in existing software systems.

With answers to those questions, the CTFS has the potential to become a standard method in STEM tomography, at least for specimen of a thickness range around 1 μm. However, the research presented in thesis is but a first step to a complete understanding of the CTFS and more questions arise as explained in the next Chapter.

Chapter 6

Future Work

6.1 Applicability and Electron Dose Restrictions

The method CTFS has so far been demonstrated on one experimental sample, consisting of gold-stained biological tissue in the thickness range of 1 μm. The method resulted in improved axial elongation, but at increased computational cost and increased electron dose. While for many applications trading almost arbitrary amounts of computational effort to increase the resolution is acceptable, the increased electron dose is a fundamental concern for dose sensitive, biological specimen.

In the introduced image acquisition scheme, for every tilt direction, a focal series of 20 images was recorded, increasing expose by the same factor. This additional electron dose was partially compensated by an increased tilt increment. The specimen was rotated in increments of 5° instead of the typical 1° steps of a conventional tilt series, reducing the expose by a factor of 5. Still, assuming a constant electron dose per image, the scheme increases the electron dose by a factor of 4. In order to compensate this, images could be recorded at a reduces electron dose, necessarily resulting in noisier images and a decreased SNR. However, it is currently unclear how the noise in the input projections of a CTFS influences the noise in the reconstructed tomogram, compared to the same process in a conventional tilt series. In other words, there is the hypothesis that increased noise in the input projections is averaged over the increased number of images in a tilt- and focal series and results in the same SNR in the final tomogram as could be expected from a tilt series of the same total electron dose. This hypothesis remains to be examined theoretically and experimentally.

6.2 Regularization and A-priori Information

The method TF-ART, introduced in (Dahmen et al., 2014a), is a typical iterative reconstruction technique derived from Kaczmarz method. In the version with the matched back projection, it is known that it converges to the minimum of the L2-norm of the residual. However, the recording scheme CTFS can be combined with a wide range of a a-priori and regularization techniques. Expectation maximization (Dempster et al., 1977) and total variation minimization (Yan & Vese, 2011) could be used in combination with a CTFS to exploit the assumption of a sparse gradient, which might work well for the case of gold nanoparticles. Similarly, the method DART (Batenburg & Sijbers, 2007) could be used to exploit the assumption that the sample consists of relatively few, known materials. Generally it can be said that the CTFS acquisition scheme can be combined almost arbitrarily with a-priori or regularization techniques, which opens a wide field of research as none of these combinations have been investigated so far.

6.3 Towards Beam Blurring Correction

The STEM transform presented in this thesis uses a double cone as a model for the electron beam. While this model represents a clear improvement over the line model typically used in tomographic reconstruction, it is still drastically simplified compared to models of the STEM probe shape (Lupini & de Jonge, 2011; Demers, Ramachandra, Drouin, & de Jonge, 2012) and electron behavior used for purposes other than tomographic reconstruction. One example for such a context is the numeric simulation of the imaging process of STEM microscopes, implemented in the CASINO software (Drouin et al., 2007; Demers et al., 2011). The software uses a Monte-Carlo approach to simulate electron trajectories described by discrete elastic scattering events. The inelastic events are either approximated by a mean energy loss model (Joy & Luo, 1989) or alternatively a hybrid model for the inelastic scattering is used where plasmon and binary electron-electron scattering events are treated as discrete events. Either way, the elastic scattering angle is determined from a random number and from tabulated cross section values from the ELSE-PA cross section software (Salvat, Jablonski, & Powell, 2005). As the physical model used in the CASINO software has been experimentally shown the be very accurate (Demers et al., 2011; Poirier-Demers, Demers, Drouin, & de Jonge, 2011), one could argue that an ideal reconstruction algorithm should use the same model for the forward projection.

However, there are two obstacles on the way towards this seemingly straight-forward approach. First, the execution of a forward projection based on a low-level physical simulation as realized in the CASINO software is prohibitively slow for an iterative reconstruction algorithm that typically needs to execute thousands or tens of thousands forward projections to find a solution. This situation might very well be overcome by carefully optimizing the implementation using approximations, pruning computations that are not relevant for a given detector (for example electrons below a given threshold of kinetic energy can be discarded if an energy filter is simulated) and careful technical optimization for a target HPC platform.

The second obstacle is more fundamental in nature. The more complex model of the electron beam mentioned before considers electron-matter interactions. This means, in this model the electron beam shape is a function of the specimen, which is described not only by its density but also by its atomic number for every point in space. In the case of a simulation of the imaging process, those values are freely available as the specimen is entered manually via a computer aided design file or comparable. In the case of a tomographic reconstruction, the specimen function is the searched for solution, i.e. not available by definition. For further considerations, let the spectroscopic specimen function h_S be fined as

$$h_S(u) \coloneqq (h, h_Z)(u). \tag{6.1}$$

Hereby, h is the specimen density function as used before and $h_Z : \mathbb{R}^4 \to \mathbb{R}$ is a function that relates a point in space to a spectrum, i.e. maps a point u and an energy loss ΔE to an intensity. As the beam blurring corrected probe shape is a function of the spectroscopic specimen function, any implementation of beam blurring corrected tomography will also have to solve the problem of spectroscopic tomography, i.e. solving for h_S.

Spectroscopic Tomography

In the context of STEM microscopy, spectroscopic images b_S can be acquired using electron energy loss spectroscopy (EELS) (Egerton, 2009; Varela et al., 2009) or energy-dispersive X-Ray spectroscopy (EDX) (Goldstein et al., 2003). Both methods have some appeal, because both EELS and EDX signals can be recorded in addition to HAADF-STEM signals with no addition electron dose. Tomographic reconstruction of EELS images assuming a line model for the electron beam can be achieved using the SIRT implementation presented in

IMOD and sparse spectrum representations acquired using principal component analysis (Jarausch, Thomas, Leonard, Twesten, & Booth, 2009; Yedra et al., 2012). Tomography using STEM-EDX has been presented (Kotula, Brewer, Michael, & Giannuzzi, 2007; Lepinay, Lorut, Pantel, & Epicier, 2013) using basically the same algorithmic approach. Either of the two approaches can could be used as basis for an iterative beam blurring correction as explained in the next section.

Iterative Beam Blurring Correction

Assuming the presence of discrete spectroscopic images b_S, one can now imagine an iterative reconstruction algorithm with beam blurring correction. The density function h_S is replaced by its discrete approximation x_S using some basis functions. The algorithm solves the problem $A_S x_S = b_S$ where A_S is the matrix representation of the probe function using successively better approximations $\bar{A}_S$ of A_S. Starting from an initial guess x_S^0 of the specimen, the probe is approximated assuming the current specimen guess as:

$$\bar{A}_S^n \coloneqq A_S(x_S^{n-1}) \tag{6.2}$$

such that the n^{th} approximation of the specimen can be computed by solving

$$\bar{A}_S^n x_S^n = b_S \tag{6.3}$$

for x_S^n. It seems intuitively quite possible that this method converges to x_S, thereby solving the spectroscopic reconstruction problem and simultaneously providing a means for beam blurring correction. However, the functions $A_S(x)$ is quite complex as it involves the electron beam simulation described above. Actually proving that the proposed method converges within any reasonable metric will therefore likely be challenging.

Thanks and Acknowledgments

Acknowledgments

The contributions concerning the recording scheme CTFS, particularly the sample preparation and image aquisition, was performed in cooperation with Prof. Dr. Niels de Jonge, Dr. Andrew Lupini, Dr. Jean-Pierre Baudoin and Dr. Christian Kübel.

The mathematical analysis of the STEM transform, particularly the derivation of the adjoint and the proof of the relation to the ray transform was conducted in cooperation with Dr. Holger Kohr.

The Ettention software package was developed in a collective effort of the team "Computational 3D Imaging" of the DFKI GmbH, the Max-Planck Institute for Informatics and EYEN SE, with contributions from Lukas Marsalek, Nico Marniok, Beata Turonova, Stefan Nickels, Sviat Bogatchev, Patrick Trampert, and me.

Research has been partially funded within the European Research Project NOTOX (FP7-267038) and the DFG grant IMCL (AOBJ: 600875). We thank the DFKI GmbH and Eyen SE for additional funding and for providing the necessary infrastructure.

We thank Eduard Arzt for support through the Leibniz Institute for New Materials. Electron microscopy was performed at the SHaRE user facility at Oak Ridge National Laboratory, sponsored by the Office of Basic Energy Sciences, US Department of Energy, and at the Karlsruhe Nano Micro Facility, a Helmholtz research infrastructure at the Karlsruhe Institute of Technology. This research was supported by the US Department of Energy, Basic Energy Sciences, Material Sciences and Engineering Division, and by NIH grant R01-GM081801.

Danksagung

Ich danke meinem Doktorvater, Prof. Dr. Philipp Slusallek, der mir mit seinem Wissen aus der Computergrafik stets zur Seite stand, mich weit über das übliche Maß hinaus förderte und mich motivierte, immer einen Schritt weiter zu denken und ein bischen tiefer zu graben.

Ich danke Prof. Dr. Niels de Jonge, ohne dessen Wissen über Mikroskopie diese Arbeit in der vorliegenden Form nie zustande gekommen wäre. Sähe die Prüfungsordnung einen zweiten Doktorvater vor, so wäre er dies.

Ich danke Herrn Dr. Holger Kohr für die phantastische Zusammenarbeit, Herrn Dr. Andrew Lupini und Herrn Lukas Marsalek für wertvolle Diskussionen und meiner Frau, Dr. Susanne Kirsch-Dahmen, für die Korrektur dieser Arbeit sowie unglaubliche Unterstützung in einer anstrengenden Zeit.

References

Aarle, W. V., van der Maar, S., Batenburg, K. J., & Sijbers, J. (2008). Computed tomography on all scales using the astra toolbox. In *Liege image days 2008: Medical imaging.*

Abbe, E. (2004). Beiträge zur Theorie des Mikroskops und der mikroskopischen Wahrnehmung. *SPIE milestone Ser.*, *178*, 12–24.

Abrishami, V., Vargas, J., Li, X., Cheng, Y., Marabini, R., Sorzano, C. O. S., & Carazo, J. M. (2015). Alignment of direct detection device micrographs using a robust Optical Flow approach. *J. Struct. Biol.*, *189*(3), 163–76.

Agulleiro, J. I., & Fernandez, J. J. (2011). Fast tomographic reconstruction on multicore computers. *Bioinformatics*, *27*(4), 582–3.

Agulleiro, J. I., Garzon, E. M., Garcia, I., & Fernandez, J. J. (2010). Vectorization with SIMD extensions speeds up reconstruction in electron tomography. *J. Struct. Biol.*, *170*(3), 570–5.

Amat, F., Moussavi, F., Comolli, L. R., Elidan, G., Downing, K. H., & Horowitz, M. (2008). Markov random field based automatic image alignment for electron tomography. *J. Struct. Biol.*, *161*(3), 260–75.

AMD. (2013). *AMD Math Libraries OpenCL Fast Fourier Transforms (FFTs) clAmdFft.* Retrieved from `http://developer.amd.com/tools-and-sdks/opencl-zone/amd-accelerated-parallel-processing-math-libraries/`

Andersen, A., & Kak, A. (1984). Simultaneous algebraic reconstruction technique (SART): a superior implementation of the ART algorithm. *Ultrason. Imaging*, *6*, 81–94.

Batenburg, K., & Sijbers, J. (2007). DART: a fast heuristic algebraic reconstruction algorithm for discrete tomography. *Image Process. 2007. ICIP 2007.*, *4*(2), 133–136.

Baudoin, J.-P., Jerome, W. G., Kübel, C., & de Jonge, N. (2013). Whole-Cell Analysis of Low-Density Lipoprotein Uptake by Macrophages Using STEM Tomography. *PLoS One*, *8*(1), e55022.

Baudoin, J.-P., Jinschek, J. R., Boothroyd, C. B., Dunin-Borkowski, R. E., & de Jonge, N. (2013). Chromatic aberration-corrected tilt series transmission electron microscopy of nanoparticles in a whole mount macrophage cell. *Microsc. Microanal.*, *19*(4), 814–20.

Beck, A., & Teboulle, M. (2009). Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Trans. Image Process.*, *18*(11), 2419–34.

Behan, G., Cosgriff, E. C., Kirkland, A. I., & Nellist, P. D. (2009). Three-dimensional imaging by optical sectioning in the aberration-corrected scanning transmission electron microscope. *Phil. Trans. A, Math. Phys. Eng. Sci.*, *367*(1903), 3825–3844.

Berriman, J., Bryan, R. K., Freeman, R., & Leonard, K. R. (1984). Methods for specimen thickness determination in electron microscopy. *Ultramicroscopy*, *13*, 351–364.

Bilbao-Castro, J. R., Marabini, R., Sorzano, C. O. S., Garcia, I., Carazo, J. M., & Fernandez, J. J. (2009). Exploiting desktop supercomputing for three-dimensional electron microscopy reconstructions using ART with blobs. *J. Struct. Biol.*, *165*(1), 19–26.

Bleloch, A., & Lupini, A. (2004). Imaging at the picoscale. *Mater. Today*, *7*, 42–48.

Borisevich, A. Y., Lupini, A. R., & Pennycook, S. J. (2006). Depth sectioning with the aberration-corrected scanning transmission electron microscope. *Proc. Natl. Acad. Sci. U. S. A.*, *103*(9), 3044–8.

Born, M., & Wolf, E. (1997). *Principles of optics.* Cambridge University Press.

Bracewell, R. N., & Riddle, A. C. (1967). *Inversion of Fan-Beam Scans in Radio Astronomy* (Vol. 150).

Bresenham, J. E. (1965). Algorithm for computer control of a digital plotter. *IBM Syst. J.*, *4*(1), 25–30.

Cao, M., Takaoka, A., Zhang, H.-B., & Nishi, R. (2011). An automatic method of detecting and tracking fiducial markers for alignment in electron tomography. *J. Electron Microsc.*, *60*(1), 39–46.

Castano Diez, D., Mueller, H., & Frangakis, A. S. (2007). Implementation and performance evaluation of reconstruction algorithms on graphics processors. *J. Struct. Biol.*, *157*(1), 288–95.

Censor, Y. (1990). On Variable Block Algebraic Reconstruction Techniques. In H. A. Dold, Z. B. Eckmann, & G. F. Takens (Eds.), *Math. methods tomogr.* (pp. 133–140). Springer.

Coene, W., Thust, A., Op de Beeck, M., & Van Dyck, D. (1996). Maximum-likelihood method for focus-variation image reconstruction in high resolution transmission electron microscopy. *Ultramicroscopy*, *64*(1-4), 109–135.

Cook, R. L. (1986). Stochastic sampling in computer graphics. *Trans. Graph.*, *5*(1), 51–72.

Dahmen, T., Baudoin, J.-P., Lupini, A. R., Kübel, C., Slusallek, P., & de Jonge, N. (2014a). Combined scanning transmission electron microscopy tilt- and focal series. *Microsc. Microanal.*, *20*(2), 548–60.

Dahmen, T., Baudoin, J.-P., Lupini, A. R., Kübel, C., Slusallek, P., & de Jonge, N. (2014b). Combined tilt- and focal series scanning transmission electron microscopy: TFS 3D STEM. In *18th int. microsc. congr. proc.* Prague.

Dahmen, T., Baudoin, J.-P., Lupini, A. R., Kübel, C., Slusallek, P., & de Jonge, N. (2014c). TFS: Combined Tilt- and Focal Series Scanning Transmission Electron Microscopy. In *Microsc. microanal.* (Vol. 20, pp. 786–787). Hartford.

Dahmen, T., Kohr, H., de Jonge, N., & Slusallek, P. (2015). Matched Backprojection Operator for Combined Scanning Transmission Electron Microscopy Tilt- and Focal Series. *under review Microsc. Microanal.*.

Dahmen, T., Marsalek, L., Turonova, B., Marniok, N., Bogatchev, S., Trampert, P., & Slusallek, P. (2015). The Ettention Software Package. *under review Ultramicroscopy*.

de Jonge, N., Sougrat, R., Northan, B. M., & Pennycook, S. J. (2010). Three-dimensional scanning transmission electron microscopy of biological specimens. *Micros. Microanal.*, *16*(1), 54–63.

Demers, H., Poirier-Demers, N., Couture, A. R., Joly, D., Guilmain, M., de Jonge, N., & Drouin, D. (2011). Three-dimensional electron microscopy simulation with the CASINO Monte Carlo software. *Scanning*, *33*(3), 135–46.

Demers, H., Ramachandra, R., Drouin, D., & de Jonge, N. (2012). The probe profile and lateral resolution of scanning transmission electron microscopy of thick specimens. *Microsc. Microanal.*, *18*(3), 582–90.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B*, *39*(1), 1–38.

Diebolder, C. A., et al. (2014). Complement is activated by IgG hexamers assembled at the cell surface. *Science*, *343*(6176), 1260–3.

Donoho, D. L. (2006). For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution. *Commun. Pure Appl. Math.*, *59*(6), 797–829.

Drouin, D., Couture, A. R., Joly, D., Tastet, X., Aimez, V., & Gauvin, R. (2007). CASINO V2.42: a fast and easy-to-use modeling tool for scanning electron microscopy and microanalysis users. *Scanning*, *29*(3), 92–101.

Dukes, M. J., Ramachandra, R., Baudoin, J.-P., Gray Jerome, W., & de Jonge, N. (2011). Three-dimensional locations of gold-labeled proteins in a whole mount eukaryotic cell obtained with 3nm precision using aberration-corrected scanning transmission electron microscopy. *J. Struct. Biol.*, *174*(3), 552–562.

Egerton, R. F. (2009). Electron energy-loss spectroscopy in the TEM. *Reports Prog. Phys.*, *72*(1), 016502.

Elfving, T., Hansen, P. C., & Nikazad, T. (2014). Semi-convergence properties of Kaczmarz's method. *Inverse Probl.*, *30*(5), 055007.

Engel, K., Hadwiger, M., Kniss, J. M., Rezk-Salama, C., & Weiskopf, D. (2006). *Real-Time Volume Graphics.* Wellesley, MA: A K Peters, Ltd.

Fernandez, J. J. (2012). Computational methods for electron tomography. *Micron*, *43*(10), 1010–1030.

Frank, J. (2006). *Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State.* Oxford University Press.

Frigo, S. P., Levine, Z. H., & Zaluzec, N. J. (2002). Submicron imaging of buried integrated circuit structures using scanning confocal electron microscopy. *Appl. Phys. Lett.*, *81*(11), 2112.

Gilbert, P. (1972a). Iterative methods for the three-dimensional reconstruction of an object from projections. *J. Theor. Biol.*, *36*(1), 105–117.

Gilbert, P. (1972b). The Reconstruction of a Three-Dimensional Structure from Projections and Its Application to Electron Microscopy. II. Direct Methods. *Proc. R. Soc. B Biol. Sci.*, *182*(1066), 89–102.

Goldstein, J., et al. (2003). *Scanning Electron Microscopy and X-ray Microanalysis* (3rd ed.). Springer.

Gordon, R., Bender, R., & Herman, G. T. (1970). Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and x-ray photography. *J. Theor. Biol.*, *29*(3), 471–481.

Goris, B., Van den Broek, W., Batenburg, K. J., Heidari Mezerji, H., & Bals, S. (2012). Electron tomography based on a total variation minimization reconstruction technique. *Ultramicroscopy*, *113*, 120–130.

Guan, H., & Gordon, R. (1994). A projection access order for speedy convergence of ART (algebraic reconstruction technique): a multilevel scheme for computed tomography. *Phys. Med. Biol.*, *39*(11), 2005–2022.

Haider, M., Uhlemann, S., & Zach, J. (2000). Upper limits for the residual aberrations of a high-resolution aberration-corrected STEM. *Ultramicroscopy*, *81*, 163–175.

Hell, S. W. (2007). Far-field optical nanoscopy. *Science*, *316*(5828), 1153–1158.

Heymann, J. B., Cardone, G., Winkler, D. C., & Steven, A. C. (2008). Computational resources for cryo-electron tomography in Bsoft. *J. Struct. Biol.*, *161*(3), 232–42.

Hoenger, A., & McIntosh, J. R. (2009). Probing the macromolecular organization of cells by electron tomography. *Curr. Opin. Cell Biol.*, *21*(1), 89–96.

Hohmann-Marriott, M. F., Sousa, A. A., Azari, A. A., Glushakova, S., Zhang, G., Zimmerberg, J., & Leapman, R. D. (2009). Nanoscale 3D cellular imaging by axial scanning transmission electron tomography. *Nat. Meth.*, *6*(10), 729–731.

Hrabe, T., Chen, Y., Pfeffer, S., Cuellar, L. K., Mangold, A.-V., & Förster, F. (2012). PyTom: a python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis. *J. Struct. Biol.*, *178*(2), 177–88.

Hsieh, W.-K., Chen, F.-R., Kai, J.-J., & Kirkland, a. I. (2004). Resolution extension and exit wave reconstruction in complex HREM. *Ultramicroscopy*, *98*(2-4), 99–114.

Hüe, F., Rodenburg, J. M., Maiden, A. M., & Midgley, P. A. (2011). Extended ptychography in the transmission electron microscope: possibilities and limitations. *Ultramicroscopy*, *111*(8), 1117–23.

Hüe, F., Rodenburg, J. M., Maiden, A. M., Sweeney, F., & Midgley, P. A. (2010). Wave-front phase retrieval in transmission electron microscopy via ptychography. *Phys. Rev. B*, *82*(12), 121415.

Humphry, M. J., Kraus, B., Hurst, A. C., Maiden, A. M., & Rodenburg, J. M. (2012). Ptychographic electron microscopy using high-angle darkfield scattering for sub-nanometre resolution imaging. *Nat. Commun.*, *3*, 730.

Intaraprasonk, V., Xin, H. L., & Muller, D. A. (2008). Analytic derivation of optimal imaging conditions for incoherent imaging in aberration-corrected electron microscopes. *Ultramicroscopy*, *108*(11), 1454–66.

Intel Corp. (2014a). *Intel VTune Amplifier User's Guide.* `https://software.intel.com/en-us/node/529797`.

Intel Corp. (2014b). *OpenCL Design and Programming Guide for the Intel Xeon Phi Coprocessor.* `https://software.intel.com/en-us/articles/opencl-design-and-programming-guide-for-the-intel-xeon-phi-coprocessor`.

Jarausch, K., Thomas, P., Leonard, D. N., Twesten, R., & Booth, C. R. (2009). Four-dimensional STEM-EELS: enabling nano-scale chemical tomography. *Ultramicroscopy*, *109*(4), 326–37.

Jerome, W. G., Cox, B. E., Griffin, E. E., & Ullery, J. C. (2008). Lysosomal cholesterol accumulation inhibits subsequent hydrolysis of lipoprotein cholesteryl ester. *Micros. Microanal.*, *14*(2), 138–149.

Jian, Y. D., Balcan, D. C., & Dellaert, F. (2012). Generalized subgraph preconditioners for large-scale bundle adjustment. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, *7474 LNCS*(Iccv), 131–150.

Jiang, M., & Wang, G. (2003). Convergence studies on iterative algorithms for image reconstruction. *IEEE Trans. Med. Imaging*, *22*(5), 569–79.

Joy, D. C., & Luo, S. (1989). An empirical stopping power relationship for low-energy electrons. *Scanning*, *11*(4), 176–180.

Kaczmarz, S. (1937). Angenäherte Auflösung von Systemen linearer Gleichungen. In *Bull. int. l'académie pol. des sci. lett.* (pp. 355–357).

Keck, B., Hofmann, H., Scherl, H., Kowarschik, M., & Hornegger, J. (2009). GPU-accelerated SART reconstruction using the CUDA programming environment. *Proc. SPIE, Med. Imaging Phys. Med. Imaging*, 72582B–72582B–9.

Khronos Group. (2007). *The OpenCL Specification.* `https://www.khronos.org/registry/cl/sdk/1.1/docs/man/xhtml/`.

Koster, A. J., Grimm, R., Typke, D., Hegerl, R., Stoschek, A., Walz, J., & Baumeister, W. (1997). Perspectives of molecular and cellular electron tomography. *J. Struct. Biol.*, *120*(3), 276–308.

Kotula, P., Brewer, L., Michael, J., & Giannuzzi, L. (2007). Computed Tomographic Spectral Imaging: 3D STEM-EDS Spectral Imaging. *Microsc. Microanal.*, *13*(S02).

Kourkoutis, L. F., Plitzko, J. M., & Baumeister, W. (2012). Electron Microscopy of Biological Materials at the Nanometer Scale. *Ann. Rev. Mater. Res.*, *42*(1), 33–58.

Kremer, J. R., Mastronarde, D. N., & McIntosh, J. R. (1996). Computer visualization of three-dimensional image data using IMOD. *J. Struct. Biol.*, *116*(1), 71–76.

Krivanek, O. L., Dellby, N., & Lupini, A. R. (1999). Towards sub-A electron beams. *Ultramicroscopy*, *78*, 1–11.

Landweber, L. (1951). An iteration formula for Fredholm integral equations of the first kind. *Am. J. Math.*, *73*, 615–624.

Lanzavecchia, S., Cantele, F., Bellon, P. L., Zampighi, L., Kreman, M., Wright, E., & Zampighi, G. A. (2005). Conical tomography of freeze-fracture replicas: a method for the study of integral membrane proteins inserted in phospholipid bilayers. *J. Struct. Biol.*, *149*(1), 87–98.

Lawrence, M. C. (1984). Alignment of images for threedimensional reconstruction of non-periodic objects. In *Electron Microsc. Soc. S. Afr. Proc.* (Vol. 13).

Lawrence, M. C. (1992). Least-Squares Method of Alignment Using Markers. In J. Frank (Ed.), *Electron tomography* (pp. 197–204). Springer US.

Lehmann, M., & Lichte, H. (2002). Tutorial on off-axis electron holography. *Microsc. Microanal.*, *8*(6), 447–66.

Lepinay, K., Lorut, F., Pantel, R., & Epicier, T. (2013). Chemical 3D tomography of 28nm high K metal gate transistor: STEM XEDS experimental method and results. *Micron*, *47*, 43–9.

Leung, B. O., & Chou, K. C. (2011). Review of super-resolution fluorescence microscopy for biology. *Appl. Spectrosc.*, *65*(9), 967–80.

Levoy, M. (1990). Efficient ray tracing of volume data. *ACM Trans. Graph.*, *9*(3), 245–261.

Ludtke, S. J., Baldwin, P. R., & Chiu, W. (1999). EMAN: semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.*, *128*(1), 82–97.

Lupini, A. R., & de Jonge, N. (2011). The three-dimensional point spread function of aberration-corrected scanning transmission electron microscopy. *Micros. Microanal.*, *17*(5), 817–826.

Maiden, A. M., Humphry, M. J., & Rodenburg, J. M. (2012). Ptychographic transmission microscopy in three dimensions using a multi-slice approach. *J. Opt. Soc. Am. A. Opt. Image Sci. Vis.*, *29*(8), 1606–14.

Maisl, M., Porsch, F., & Schorr, C. (2010). Computed Laminography for X-ray Inspection of Lightweight Constructions. In *2nd int. symp. ndt aerosp.* (pp. 2–8).

Marabini, R., Herman, G. T., & Carazo, J. M. (1998). 3D reconstruction in electron microscopy using ART with smooth spherically symmetric volume elements (blobs). *Ultramicroscopy*, *72*(1-2), 53–65.

Mastronarde, D. N. (1997). Dual-axis tomography: an approach with alignment methods that preserve resolution. *J. Struct. Biol.*, *120*(3), 343–52.

Messaoudii, C., Boudier, T., Sanchez Sorzano, C. O., & Marco, S. (2007). TomoJ: tomography software for three-dimensional reconstruction in transmission electron microscopy. *BMC Bioinformatics*, *8*, 288.

Meyer-Ilse, W., et al. (2001). High resolution protein localization using soft X-ray microscopy. *J. Microsc.*, *201*(Pt 3), 395–403.

Mueller, K., Yagel, R., & Wheller, J. J. (1999). Anti-aliased three-dimensional cone-beam reconstruction of low-contrast objects with algebraic methods. *Med. Imaging, IEEE*, *18*(6), 519–37.

Natterer, F., & Wübbeling, F. (2001). *Mathematical Methods in Image Reconstruction.* SIAM.

Nellist, P. D., McCallum, B. C., & Rodenburg, J. M. (1995). Resolution beyond the 'information limit' in transmission electron microscopy. *Nature*, *374*(6523), 630–632.

Norton, J. S. (1985). Iterative reconstruction algorithms: convergence as a function of spatial frequency. *J. Opt. Soc. Am. A*, *2*(1), 6–13.

Olins, D. E., Olins, A. L., Levy, H. A., Durfee, R. C., Margle, S. M., Tinnel, E. P., & Dover, S. D. (1983). Electron microscope tomography: transcription in three dimensions. *Science*, *220*, 498–500.

Palenstijn, W. J., Batenburg, K. J., & Sijbers, J. (2011). Performance improvements for iterative electron tomography reconstruction using graphics processing units (GPUs). *J. Struct. Biol.*, *176*(2), 250–3.

Patwardhan, A., et al. (2012). Data management challenges in three-dimensional EM. *Nat. Struct. Mol. Biol.*, *19*(12), 1203–7.

Patwardhan, A., et al. (2014). A 3D cellular context for the macromolecular world. *Nat. Struct. Mol. Biol.*, *21*(10), 841–845.

Penczek, P., Marko, M., Buttle, K., & Frank, J. (1995). Double-tilt electron tomography. *Ultramicroscopy*, *60*(3), 393–410.

Penczek, P., Zhu, J., & Schröder, R. (1997). Three dimensional reconstruction with contrast transfer compensation from defocus series. *Scanning Microsc.*, *11*(518), 147–154.

Pennycook, S. J., & Nellist, P. D. (2011). *Scanning Transmission Electron Microscopy: Imaging and Analysis.* Springer.

Poirier-Demers, N., Demers, H., Drouin, D., & de Jonge, N. (2011). Simulating STEM Imaging of Nanoparticles in Micrometers-Thick Substrates. *Microsc. Microanal.*, *17*(S2), 980–981.

Qu, G., Wang, C., & Jiang, M. (2009). Necessary and sufficient convergence conditions for algebraic image reconstruction algorithms. *IEEE Trans. Image Process.*, *18*(2), 435–40.

Radermacher, M. (1992). Weighted Back-Projection Methods. In J. Frank (Ed.), *Electron tomography* (pp. 91–115). Springer.

Ramachandra, R., & de Jonge, N. (2012). Optimized deconvolution for maximum axial resolution in three-dimensional aberration-corrected scanning transmission electron microscopy. *Microsc. Microanal.*, *18*(1), 218–28.

Ramachandran, G. N., & Lakshminarayanan, A. V. (1971). Three-dimensional reconstruction from radiographs and electron micrographs: application of convolutions instead of Fourier transforms. *Proc. Natl. Acad. Sci. U. S. A.*, *68*, 2236–2240.

Reimer, L. (1998). *Scanning Electron Microscopy: Physics of Image Formation and Microanalysis*. Springer.

Ress, D. B., Harlow, M. L., Marshall, R. M., & McMahan, U. J. (2004). Methods for generating high-resolution structural models from electron microscope tomography data. *Structure*, *12*(10), 1763–74.

Ring, E. A., Peckys, D. B., Dukes, M. J., Baudoin, J. P., & de Jonge, N. (2011). Silicon nitride windows for electron microscopy of whole cells. *J. Microsc.*, *243*(3), 273–283.

Robert, C. P., & Casella, G. (2005). *Monte Carlo Statistical Methods*. Springer.

Rodriguez, M. B., Gobbetti, E., Guitian, J. A. I., Makhinya, M., Marton, F., Pajarola, R., & Suter, S. K. (2013). A Survey of Compressed GPU-Based Direct Volume Rendering. *Eurographics State-of-the-art Report*, 117–136.

Rudin, L. I., Osher, S., & Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Phys. D Nonlinear Phenom.*, *60*(1-4), 259–268.

Rudin, W. (1987). *Real and complex analysis*. McGraw-Hill.

Russ, J. C. (2006). *The image processing handbook* (5th ed.). CRC Press.

Salvat, F., Jablonski, A., & Powell, C. J. (2005). elsepa - Dirac partial-wave calculation of elastic scattering of electrons and positrons by atoms, positive ions and molecules. *Comput. Phys. Commun.*, *165*(2), 157–190.

Scheres, S. H. W. (2012). RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.*, *180*(3), 519–30.

Scherl, H., Keck, B., Kowarschik, M., & Hornegger, J. (2007). Fast GPU-Based CT Reconstruction using the Common Unified Device Architecture (CUDA). *2007 IEEE Nucl. Sci. Symp. Conf. Rec.*, 4464–4466.

Schoenmakers, R. H. M., Perquin, R. A., Fliervoet, T. F., Voorhout, W., & Schirmacher, H. (2005). New Software for High Resolution, High Throughput Electron Tomography. *Microscopy and Analysis*, *19*(4), 5–7.

Shaikh, T. R., Gao, H., Baxter, W. T., Asturias, F. J., Boisset, N., Leith, A., & Frank, J. (2008). SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. *Nat. Protoc.*, *3*(12), 1941–74.

Smyth, M. S., & Martin, J. H. (2000). x ray crystallography. *Mol. Pathol.*, *53*, 8–14.

Sorzano, C. O. S., et al. (2009). Marker-free image registration of electron tomography tilt-series. *BMC Bioinformatics*, *10*, 124.

Tang, G., Peng, L., Baldwin, P. R., Mann, D. S., Jiang, W., Rees, I., & Ludtke, S. J. (2007). EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.*, *157*(1), 38–46.

Van Dyck, D. (2010). Wave reconstruction in TEM using a variable phase plate. *Ultramicroscopy*, *110*(5), 571–572.

Van Dyck, D., Jinschek, J. R., & Chen, F.-R. (2012). 'Big Bang' tomography as a new route to atomic-resolution electron tomography. *Nature*, *486*(7402), 243–6.

van Heel, M., Harauz, G., Orlova, E. V., Schmidt, R., & Schatz, M. (1996). A new generation of the IMAGIC image processing system. *J. Struct. Biol.*, *116*(1), 17–24.

van Heel, M., & Keegstra, W. (1981). IMAGIC: A fast, flexible and friendly image analysis software system. *Ultramicroscopy*, *7*(2), 113–129.

Varela, M., et al. (2009). Atomic-resolution imaging of oxidation states in manganites. *Phys. Rev. B*, *79*(8), 085117.

Vazquez, F., Garzon, E. M., & Fernandez, J. J. (2011). Matrix Implementation of Simultaneous Iterative Reconstruction Technique (SIRT) on GPUs. *Comput. J.*, *54*(11), 1861–1868.

Venkatakrishnan, S. V., Drummy, L. F., De Graef, M., Simmons, J. P., & Bouman, C. A. (2013). Model based iterative reconstruction for Bright Field electron tomography. *Proc. of SPIE-IS T Electronic Imaging*, *8657*.

Voortman, L. M., Stallinga, S., Schoenmakers, R. H. M., van Vliet, L. J., & Rieger, B. (2011). A fast algorithm for computing and correcting the CTF for tilted, thick specimens in TEM. *Ultramicroscopy*, *111*(8), 1029–36.

Wang, J., Zheng, Y., & Member, S. (2007). On the Convergence of Generalized Simultaneous Iterative Reconstruction Algorithms. *IEEE Trans. Image Process.*, *16*(1), 1–6.

Watt, Allan. (2000). *3D Computer Graphics* (3rd ed.). Addison Wesley.

Whitted, T. (1980). An improved illumination model for shaded display. *Comm. ACM*, *23*(6), 343–349.

Williams, D. B., & Carter, C. B. (2009). *Transmission Electron Microscopy.* Springer.

Winkler, H. (2007). 3D reconstruction and processing of volumetric data in cryo-electron tomography. *J. Struct. Biol.*, *157*(1), 126–37.

Winkler, H., & Taylor, K. A. (2006). Accurate marker-free alignment with simultaneous geometry determination and reconstruction of tilt series in electron tomography. *Ultramicroscopy*, *106*, 240–254.

Wolf, D., Lubk, A., & Lichte, H. (2014). Weighted simultaneous iterative reconstruction technique for single-axis tomography. *Ultramicroscopy*, *136*, 15–25.

Xu, F., & Mueller, K. (2005). Accelerating popular tomographic reconstruction algorithms on commodity PC graphics hardware. *Nucl. Sci. IEEE Trans.*, *52*(3), 654–663.

Xu, W., et al. (2010). High-performance iterative electron tomography reconstruction with long-object compensation using graphics processing units (GPUs). *J. Struct. Biol.*, *171*(2), 142–153.

Yan, M. (2010). Convergence Analysis of SART by Bregman Iteration and Dual Gradient Descent. *Proc. SIAM Conference on Comp. Science and Engineering*, 1–15.

Yan, M., & Vese, L. A. (2011). Expectation maximization and total variation-based model for computed tomography reconstruction from undersampled data. *Proc. Spie 7691 Med. Imaging 2011 Phys. Med. Imaging*, *7961*.

Yedra, L., et al. (2012). EEL spectroscopic tomography: towards a new dimension in nanomaterials analysis. *Ultramicroscopy*, *122*, 12–8.

Zeng, G. L., & Gullberg, G. T. (2000). Unmatched projector/backprojector pairs in an iterative reconstruction algorithm. *IEEE Trans. Med. Imaging*, *19*(5), 548–55.

Zheng, S. Q., Keszthelyi, B., Branlund, E., Lyle, J. M., Braunfeld, M. B., Sedat, J. W., & Agard, D. A. (2007). UCSF tomography: an integrated software suite for real-time electron microscopic tomographic data collection, alignment, and reconstruction. *J. Struct. Biol.*, *157*(1), 138–47.

List of Abbreviations

1D	One-dimensional
2D	Two-dimensional
3D	Three-dimensional
API	Application programming interface
ART	Algebraic reconstruction technique
CCD	Charge-coupled device
CPU	Central processing unit
CT	Computed tomography
CTEM	Conventional transmission electron microscope
CTFS	Combined tilt- and focal series
DART	Discrete algebraic reconstruction technique
DOF	Depth of field
EDX	Energy-dispersive X-Ray spectroscopy
EELS	Electron energy loss spectroscopy
FFT	Fast Fourier transform
FWHM	Full-width-at-half-maximum
GB	Gigabyte
GPU	Graphics processing unit
HAADF	High angle annular dark field
HPC	High performance computing
MB	Megabyte
mrad	Miliradians
ms	Miliseconds
nm	Nanometer
PCI	Peripheral component interconnect
PSF	Point spread function
RMSE	Root means square error
s	Seconds
SART	Simultaneous algebraic reconstruction technique
SIMD	Single instruction multiple data

SIRT	Sequential iterative reconstruction technique
SNR	Signal-to-noise ratio
STEM	Scanning transmission electron microscopy
TEM	Transmission electron microscope
TF-ART	Tilt-focal algebraic reconstruction technique
TVM	Total variation minimization
WBP	Weighted back projection

List of Symbols

Overview of symbols and notations used in this thesis. Rs: Real space, Fs: Fourier space. Table from (Dahmen, Kohr, et al., 2015).

domain	semantic	symbol	remark
vectors (Rs)	coordinates in volume	$\boldsymbol{u}$	$\boldsymbol{u} = (x, y, z)$, rotated: $\boldsymbol{u}' = (x', y', z')$
	focal spot position	$\boldsymbol{v}$	vertex of the double cone
	first two components of $\boldsymbol{u}$	$\bar{\boldsymbol{u}}$	
	beam axis	$\boldsymbol{\theta}$	unit length ($\lvert\boldsymbol{\theta}\rvert = 1$)
	other direction in cone	$\boldsymbol{\omega}$	unit length
	perpendicular vector	$\boldsymbol{\eta}$	such that $\boldsymbol{u}' = s\boldsymbol{\theta} + \boldsymbol{\eta}$
scalars (Rs)	disc radius	r	usually $r = \lvert z'\rvert \tan\alpha$
vectors (Fs)	spatial frequency	$\boldsymbol{\xi}$	rotated: $\boldsymbol{\xi}'$
	first two components of $\boldsymbol{\xi}$	$\bar{\boldsymbol{\xi}}$	
	perpendicular vector	$\boldsymbol{\zeta}$	such that $\boldsymbol{\xi}' = \sigma\boldsymbol{\theta} + \boldsymbol{\zeta}$
scalars (Fs)	frequency along $\boldsymbol{\theta}$	σ	
angles	beam opening semi-angle	α	
	tilt angle	β	tilt around y-axis
functions	searched-for tomogram	$f(\boldsymbol{u})$	Fourier transform: $\widehat{f}(\boldsymbol{\xi})$
	probe function (PSF)	$p(\boldsymbol{u})$	rotated: $p_{\boldsymbol{\theta}}(\boldsymbol{u}')$, Fourier transform: $\widehat{p}(\boldsymbol{\xi})$
	focal stack (data volume)	$g_{\boldsymbol{\theta}}(\boldsymbol{v})$	projection direction: $\boldsymbol{\theta}$
		$g_k(\boldsymbol{v})$	denotes $g_{\boldsymbol{\theta}}(\boldsymbol{v})$ for $\boldsymbol{\theta} = \boldsymbol{\theta}_k$
operators	forward projection operator	$\mathcal{A}_{\boldsymbol{\theta}}$	linear convolution
	back projection operator	$\mathcal{A}^*_{\boldsymbol{\theta}}$	adjoint of $\mathcal{A}_{\boldsymbol{\theta}}$
coefficients	searched-for tomogram	X	discrete approx. of $f(\boldsymbol{u})$
	focal stack	B_k	discrete approx. of $g_k(\boldsymbol{v})$

Appendix

Derivation of the Relationship Between the STEM Transform and the Parallel Ray Transform

This derivation relates to the actual computation of the transform in that it rephrases the operator as a collection of line integrals parametrized by points in a disc perpendicular to $\boldsymbol{\theta}$ at a fixed distance $d > 0$ from the origin. One starts by rewriting the integral

$$\int_{C'_\alpha} F(\boldsymbol{u}')\,\mathrm{d}\boldsymbol{u}' = \int_{\substack{\boldsymbol{\eta}\in\boldsymbol{\theta}^\perp \\ |\boldsymbol{\eta}|<d\tan\alpha}} \int_{\mathbb{R}} F\big(s\boldsymbol{\omega}(\boldsymbol{\eta})\big)\,\tau(s,\boldsymbol{\eta})\,\mathrm{d}s\,\mathrm{d}\boldsymbol{\eta} \tag{A.1}$$

for a general function $F : \mathbb{R}^3 \to \mathbb{R}$ with an appropriate weight τ and the direction vector

$$\boldsymbol{\omega}(\boldsymbol{\eta}) = \frac{d\boldsymbol{\theta} + \boldsymbol{\eta}}{\sqrt{d^2 + |\boldsymbol{\eta}|^2}}. \tag{A.2}$$

This means that for any point $\boldsymbol{0} \neq \boldsymbol{u}' \in C'_\alpha$, one determines the corresponding $\boldsymbol{\eta} \in \boldsymbol{\theta}^\perp$ such that $\boldsymbol{\omega}(\boldsymbol{\eta}) = \boldsymbol{u}'/\,|\boldsymbol{u}'|$, i.e. $\boldsymbol{u}'$ and $d\boldsymbol{\theta}+\boldsymbol{\eta}$ lie on the same line. For simplicity, the case $\boldsymbol{\theta} = \boldsymbol{e}_z$ is considered first. One determines a vector $(\bar{\boldsymbol{p}}, d)$ which is collinear with $\boldsymbol{u}$. Such a vector is given by

$$\bar{\boldsymbol{p}} = \frac{d}{u_z}\,\bar{\boldsymbol{u}} \tag{A.3}$$

since $\boldsymbol{u} = \frac{u_z}{d}\,(\bar{\boldsymbol{p}}, d)$. Setting $t = u_z/d$, this defines a mapping

$$\boldsymbol{u} = \Gamma(\bar{\boldsymbol{p}}, t) = (t\bar{\boldsymbol{p}}, td). \tag{A.4}$$

Its Jacobi matrix of derivatives is given by

$$\mathrm{D}\Gamma(\bar{\boldsymbol{p}},t) = \begin{pmatrix} t & 0 & p_1 \\ 0 & t & p_2 \\ 0 & 0 & d \end{pmatrix}, \tag{A.5}$$

such that the functional determinant in the integral reparametrization is

$$\det \mathrm{D}\Gamma(\bar{\boldsymbol{p}},t) = t^2 d. \tag{A.6}$$

Now the left-hand side of (A.1) for $\boldsymbol{\theta} = \boldsymbol{e}_z$ can be rewritten as

$$\int_{C_\alpha} F(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} = \int_{\substack{\bar{\boldsymbol{p}}\in\mathbb{R}^2 \\ |\bar{\boldsymbol{p}}|<d\tan\alpha}} \int_{\mathbb{R}} F\big(\Gamma(\bar{\boldsymbol{p}},t)\big)\, t^2 d\,\mathrm{d}t\,\mathrm{d}\bar{\boldsymbol{p}} \tag{A.7}$$

$$= \int_{\substack{\bar{\boldsymbol{p}}\in\mathbb{R}^2 \\ |\bar{\boldsymbol{p}}|<d\tan\alpha}} \int_{\mathbb{R}} F\big(t(\bar{\boldsymbol{p}},d)\big)\, t^2 d\,\mathrm{d}t\,\mathrm{d}\bar{\boldsymbol{p}}. \tag{A.8}$$

Finally, by normalizing the vector $(\bar{\boldsymbol{p}},d)$ in the argument of F with the reparametrization $t = s/\sqrt{d^2 + |\bar{\boldsymbol{p}}|^2}$, one acquires

$$\int_{C_\alpha} F(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} = \int_{\substack{\bar{\boldsymbol{p}}\in\mathbb{R}^2 \\ |\bar{\boldsymbol{p}}|<d\tan\alpha}} \int_{\mathbb{R}} F\big(s\boldsymbol{\omega}(\bar{\boldsymbol{p}})\big)\, s^2 d\,(d^2 + |\bar{\boldsymbol{p}}|^2)^{-3/2}\,\mathrm{d}s\,\mathrm{d}\bar{\boldsymbol{p}}. \tag{A.9}$$

The result for general $\boldsymbol{\theta}$ can be accomplished by rotation, resulting in

$$\int_{C'_\alpha} F(\boldsymbol{u}')\,\mathrm{d}\boldsymbol{u}' = \int_{\substack{\boldsymbol{\eta}\in\boldsymbol{\theta}^\perp \\ |\boldsymbol{\eta}|<d\tan\alpha}} \int_{\mathbb{R}} F\big(s\boldsymbol{\omega}(\boldsymbol{\eta})\big)\, s^2 d\,(d^2 + |\boldsymbol{\eta}|^2)^{-3/2}\,\mathrm{d}s\,\mathrm{d}\boldsymbol{\eta}. \tag{A.10}$$

To relate this to the STEM transform, one inserts $F(\boldsymbol{u}') = p_{\boldsymbol{\theta}}(\boldsymbol{u}')\,f(\boldsymbol{v}-\boldsymbol{u}')$, which yields

$$[\mathcal{A}_{\boldsymbol{\theta}} f](\boldsymbol{v}) = \int_{C'_\alpha} p_{\boldsymbol{\theta}}(\boldsymbol{u}')\,f(\boldsymbol{v}-\boldsymbol{u}')\,\mathrm{d}\boldsymbol{u}' \tag{A.11}$$

$$= \int_{\substack{\boldsymbol{\eta}\in\boldsymbol{\theta}^\perp \\ |\boldsymbol{\eta}|<d\tan\alpha}} \int_{\mathbb{R}} p_{\boldsymbol{\theta}}\big(s\boldsymbol{\omega}(\boldsymbol{\eta})\big)\, f\big(\boldsymbol{v} - s\boldsymbol{\omega}(\boldsymbol{\eta})\big)\, s^2 d\,(d^2 + |\boldsymbol{\eta}|^2)^{-3/2}\,\mathrm{d}s\,\mathrm{d}\boldsymbol{\eta}. \tag{A.12}$$

The value of $p_{\boldsymbol{\theta}}$ at $s\boldsymbol{\omega}(\boldsymbol{\eta})$ equals the value of the weight function w at the component of $s\boldsymbol{\omega}(\boldsymbol{\eta})$ along $\boldsymbol{\theta}$, see the definition Equation 2.23 of $p_{\boldsymbol{\theta}}$. This component is given by $sd/\sqrt{d^2 + |\boldsymbol{\eta}|^2}$, and in consequence,

$$p_{\boldsymbol{\theta}}\big(s\boldsymbol{\omega}(\boldsymbol{\eta})\big) = \frac{d^2 + |\boldsymbol{\eta}|^2}{\pi \tan^2 \alpha \, d^2 s^2}. \tag{A.13}$$

This yields

$$[\mathcal{A}_{\boldsymbol{\theta}} f](\boldsymbol{v}) = \frac{1}{\pi \tan^2 \alpha \, d} \int_{\substack{\boldsymbol{\eta} \in \boldsymbol{\theta}^\perp \\ |\boldsymbol{\eta}| < d \tan \alpha}} (d^2 + |\boldsymbol{\eta}|^2)^{-1/2} \int_{\mathbb{R}} f\big(\boldsymbol{v} - s\boldsymbol{\omega}(\boldsymbol{\eta})\big) \, \mathrm{d}s \, \mathrm{d}\boldsymbol{\eta}. \tag{A.14}$$

The inner integral can be read as a ray transform

$$\int_{\mathbb{R}} f\big(\boldsymbol{v} - s\boldsymbol{\omega}(\boldsymbol{\eta})\big) \, \mathrm{d}s = \int_{\mathbb{R}} f\big(s\boldsymbol{\omega}(\boldsymbol{\eta}) + \boldsymbol{v}\big) \, \mathrm{d}s = [\mathcal{P}_{\boldsymbol{\omega}(\boldsymbol{\eta})} f](\boldsymbol{v}). \tag{A.15}$$

Finally, the STEM transform can be rewritten as

$$[\mathcal{A}_{\boldsymbol{\theta}} f](\boldsymbol{v}) = \frac{1}{\pi d^2 \tan^2 \alpha} \int_{\substack{\boldsymbol{\eta} \in \boldsymbol{\theta}^\perp \\ |\boldsymbol{\eta}| < d \tan \alpha}} \frac{d}{\sqrt{d^2 + |\boldsymbol{\eta}|^2}} [\mathcal{P}_{\boldsymbol{\omega}(\boldsymbol{\eta})} f](\boldsymbol{v}). \tag{A.16}$$

To compute the value of the transform in the limit $\alpha \to 0$, the Lebesgue mean value theorem (W. Rudin, 1987) can be applied which states that the mean value integral over balls B_R with radii $R > 0$ centered at zero converges to the function value at zero,

$$\lim_{R \to 0} \frac{1}{|B_R|} \int_{B_R} \psi(\boldsymbol{\eta}) \, \mathrm{d}\boldsymbol{\eta} = \psi(\mathbf{0}), \tag{A.17}$$

provided that ψ is locally integrable. If f is square-integrable and does not extend infinitely along the beam, i.e. if the intersection of C'_α with the support of f is a bounded set, this assumption holds true for the function

$$\psi(\boldsymbol{\eta}) = \frac{d}{\sqrt{d^2 + |\boldsymbol{\eta}|^2}} [\mathcal{P}_{\boldsymbol{\omega}(\boldsymbol{\eta})} f](\boldsymbol{v}), \tag{A.18}$$

and using the fact that $\pi d^2 \tan^2 \alpha$ is the area of the integration set yields

$$\lim_{\alpha \to 0} \frac{1}{\pi d^2 \tan^2 \alpha} \int_{\substack{\boldsymbol{\eta} \in \boldsymbol{\theta}^\perp \\ |\boldsymbol{\eta}| < d \tan \alpha}} \frac{d}{\sqrt{d^2 + |\boldsymbol{\eta}|^2}} [\mathcal{P}_{\boldsymbol{\omega}(\boldsymbol{\eta})} f](\boldsymbol{v}) = [\mathcal{P}_{\boldsymbol{\omega}(\mathbf{0})} f](\boldsymbol{v}) = [\mathcal{P}_{\boldsymbol{\theta}} f](\boldsymbol{v}). \tag{A.19}$$