SCIENTIFIC DATA

DATA DESCRIPTOR

A cone-beam X-ray computed tomography data collection designed for machine learning

Henri Der Sarkissian[1*], Felix Lucka[1,2*], Maureen van Eijnnatten[1], Giulia Colacicco[1], Sophia Bethany Coban[1] & Kees Joost Batenburg[1,3]

Unlike previous works, this open data collection consists of X-ray cone-beam (CB) computed tomography (CT) datasets specifically designed for machine learning applications and high cone-angle artefact reduction. Forty-two walnuts were scanned with a laboratory X-ray set-up to provide not only data from a single object but from a class of objects with natural variability. For each walnut, CB projections on three different source orbits were acquired to provide CB data with different cone angles as well as being able to compute artefact-free, high-quality ground truth images from the combined data that can be used for supervised learning. We provide the complete image reconstruction pipeline: raw projection data, a description of the scanning geometry, pre-processing and reconstruction scripts using open software, and the reconstructed volumes. Due to this, the dataset can not only be used for high cone-angle artefact reduction but also for algorithm development and evaluation for other tasks, such as image reconstruction from limited or sparse-angle (low-dose) scanning, super resolution, or segmentation.

Background & Summary

Scientific context.   X-ray computed tomography (CT) is a widely used projection-based imaging modality with a broad range of clinical, scientific and industrial applications. In many of those, CT scanners use a particular projection geometry called circular *cone-beam* (*CB*). This scanning geometry typically leads to a distinct type of artefact in the image regions with a high cone angle, cf. Fig. 1. While several reconstruction or correction methods have been proposed to reduce high cone angle artefacts[1–3], they remain a crucial drawback of CBCT scanners over other scanners, which have disadvantages such as higher radiation dose or costs in return[4].

In the scientific field, there is often a clear division between computational imaging groups with a background in mathematics and computer science, which focus on enhancing CT methodology on one side and experimental imaging groups using CT as a tool to conduct their scientific studies on the other. The latter typically uses commercial CT solutions coming with proprietary software which does not give full access to the raw projection data or the details of the experimental acquisition. As a result, many mathematical and computational studies rely on artificial data simulated with varying degrees of realism. This lack of suitable experimental data is a significant hurdle for the translation of innovative research into applications.

Many important, recent CT innovations introduce machine learning techniques into the tomographic image reconstruction process[5,6], in particular deep neural networks (*deep learning*). For these approaches, realistic experimental data are not only needed for evaluation but more crucially, for constructing the method itself. Namely, the network parameters are optimized based on *training data* which consists of a large number of representative pairs of input data with the desired ideal output of the network (*ground truth*). While many large, open, bench-mark data collections meeting these criteria exist for standard applications of deep learning (e.g., MNIST[7] for the classification of handwritten digits), there are very few suitable projection datasets for deep learning for CT so far: Several open fan beam (2D) and cone beam (3D) X-ray CT datasets acquired by a laboratory set-up or with synchrotron parallel X-ray sources have been published in previous works[8–12]. Suitable clinical datasets are more difficult to acquire and distribute openly. The ultimate quality measure for clinical images is their diagnostic

[1]Centrum Wiskunde en Informatica, Computational Imaging group, Science Park 123, 1098XG, Amsterdam, The Netherlands. [2]Centre for Medical Image Computing, University College London, WC1E 6BT, London, United Kingdom. [3]Leiden University, Department of Mathematics, 2300 RA, Leiden, The Netherlands. *email: henri.dersarkissian@gmail.com; Felix.Lucka@cwi.nl
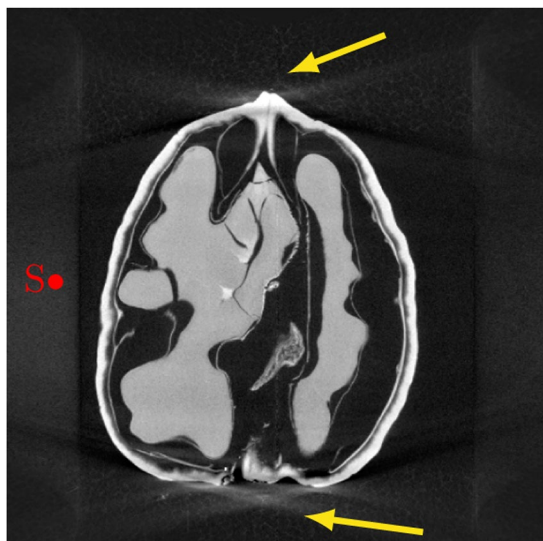
Fig. 1 Vertical slice through an FDK reconstruction of a CBCT scan of a walnut. The red dot indicates the vertical position of the source orbit and the yellow arrows point at the high cone angle artefacts.

value, which needs to be assessed by radiologists. Therefore, data are often only published as part of an image reconstruction challenge. A prominent example of this is the Mayo Clinic Low Dose CT challenge[13] consisting of 3D helical CT abdominal scans of ten cancer patients. While these datasets are extremely useful to evaluate reconstruction algorithms on a wide range of different objects and acquisition conditions, they are not suitable for machine learning as they typically contain only a single or very few scanned objects or have not been designed such that the reconstruction quality can easily be assessed in an automated way with respect to a high-quality ground truth reconstruction.

To fill this gap, we acquired a carefully designed CBCT data collection suitable for developing machine learning approaches: 42 walnuts (this choice will be discussed in the next section) were scanned with a special laboratory X-ray CBCT scanner. For each sample, CB projections were acquired on 3 different circular orbit heights. This creates different cone angles and resulting artefact pattern as well as allowing for an artefact-free, high-quality ground truth to be computed from the combined data. We provide reconstructed volumes and an open software implementation of the complete image reconstruction pipeline along with the raw projection data. Note that while 42 samples seem few compared to the training data sizes used in other deep learning applications, each sample here is a 3D object. Extracting 2D slices from these high-resolution volumes composed of 501^3 voxels gives enough data for training 2D networks that are then used to process volumes slice-by-slice, which is currently the most common approach in 3D applications[14]. While this dataset is designed to benchmark machine-learning-based correction techniques for CB artefacts, it can also be used for algorithm development and evaluation for other tomography applications, such as image reconstruction from limited or sparse-angle (low-dose) scanning, super resolution, or for image analysis tasks such as semantic segmentation.

Methods

Sample collection. A data set suitable for deep learning with convolutional layers (*convolutional neural networks*, *CNNs*) needs to be collected in a particular way. During training, the network needs to learn to recognize common spatial features and their natural variations of the class of objects that should be imaged. For this, data acquired from a sufficiently large number of representative samples is needed. Having too few samples to train on can lead to over-fitting and reduce the network's ability to generalize to unseen data. Partly inspired by[12], we decided to scan 42 walnuts: Similar to objects scanned in (pre-) clinical imaging, they contain natural inter-population variability which is advantageous compared to manufactured objects like phantoms used to calibrate scanners. They consist of a hard shell, a softer inside, air filled cavities and a variety of large-to-fine-scale details which makes them a good proxy for the human head. In addition, their size ($\approx$3 cm height) is suitable for our experimental set-up.

X-Ray tomography scanner. The scans were performed using a custom-built, highly flexible X-ray CT scanner, the FleX-ray scanner, developed by XRE nv (https://xre.be/) and located in the FleX-ray Lab at the Centrum Wiskunde & Informatica (CWI) in Amsterdam, Netherlands. The general purpose of the FleX-ray Lab is to conduct proof-of-concept experiments directly accessible to researchers in the field of mathematics and computer science. The scanner consists of a cone-beam microfocus X-ray point source (limited to 90 kV and 90 W) that projects polychromatic X-rays onto a 1536×1944 pixels, 74.8 μm^2 each, 14-bit flat panel detector (Dexella 1512NDT) and a rotation stage in-between, upon which the sample is mounted. All three components are mounted on translation stages which allow them to move independently from one another. A schematic view of the set-up with the description of possible movements is shown in Fig. 2.
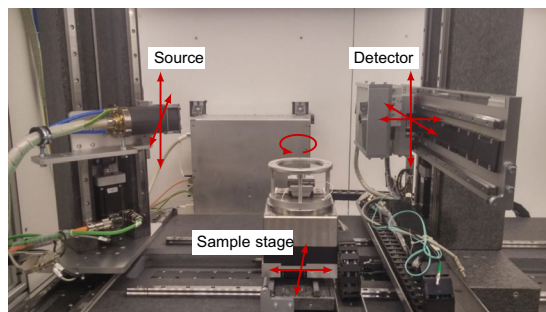
Fig. 2 FleX-ray Lab: the X-ray cone-beam tomography set-up used for the data acquisition. The arrows indicate the degrees of freedom.
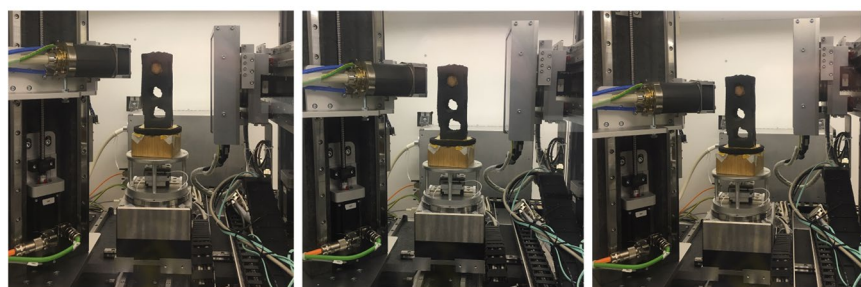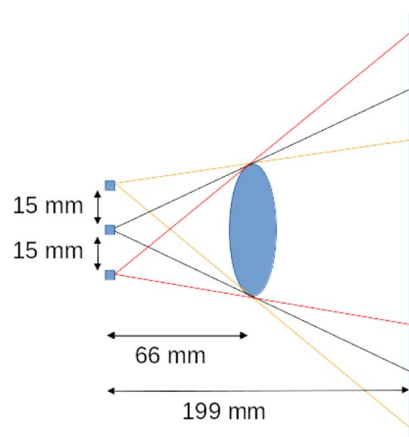

Fig. 3 Scanning geometry and trajectories for each sample. Top row: Schematic view from the side. Three full circular orbits are recorded at 3 distinct source and detector heights. The 3 squares on the left denote the source positions. Bottom row: Photographs of actual realization.

Projection geometry and acquisition parameters. Our aim was to create a data collection suitable for *supervised learning*. In supervised learning, the training data consists of pairs of input data with the desired ideal output of the network (the ground truth). A distance function (*training loss*) between ground truth and current output of the network is used to drive the optimization of the network's parameters. In our case, the input to the network may be the artefact-ridden reconstruction of a sample computed from a single orbit CBCT data set, and the ground truth could be the corresponding, high-quality, artefact-free reconstruction. We thus needed to acquire projection data from which both of these reconstructions can be computed. To obtain severe high cone angle artefacts, we needed to maximize the vertical cone-beam angle by moving the sample as close as possible to the source and choose an appropriate detector-to-object distance to maximize magnification while keeping the sample in the field of view at all times. Then, we varied the source height to collect projections from 3 circular orbits, cf. Fig. 3 (the detector height needed to be adjusted accordingly in order to fit the entire sample in every projection). In the following section, we will see that while the reconstructions from each orbit alone have different artefact pattern, combining the data from all orbits gives a high-quality reconstruction free of high cone angle artefacts.

Each walnut was embedded in a foam mount (cf. Fig. 3 bottom row). This foam is almost transparent to the X-ray beam used in our experiment. For each orbit, 1201 projections were taken during a continuous, full rotation of the sample. First and last projection were taken at the same position, leading to an angular increment of 0.3°.

Tube voltage	40 kV
Tube power	12 W
Exposure time	80 ms
Number of averages	1
Hardware binning	2×2 pixels
Effective detector pixel size	149.6 μm
Detector rows	972
Detector colums	768
Source to object distance	66 mm
Source to detector distance	199 mm
Magnification	3.016
Number of projections per orbit	1201
Angular increment	0.3°

Table 1. Summary of the acquisition parameters used.

The exposure time for each projection was 80 ms and the acquired data was binned on the fly by 2-by-2 pixel windows, i.e. each raw projection was of size 768×972 pixels. Each binned detector pixel is sized $149.6 \times 149.6 \, \mu\text{m}^2$ for a total detection field of view of $114.89 \times 145.41 \, \text{mm}^2$. During the experiment, the source voltage and power were set to 40 kV and 12 W, respectively. These values had been adjusted to ensure maximum contrast in the projection domain while avoiding detector saturation. Table 1 summarizes the acquisition parameters used.

Before every orbital scan, the source was turned off to record a projection of the detector offset count, the so-called *dark-field* image. After switching the source on again, a projection was recorded without the sample in the field of view, the so-called *flat-field* image showing the beam profile. A second flat-field was recorded after the orbital scan to correct for shadowing effects. Flat-field and dark-field images can be used to pre-process the raw photon count data for the image reconstruction as described in the next section. Examples of the projections collected for each sample are shown in Fig. 4.

Reconstructed volumes. Each projection image P consist of raw photon counts per detector pixel that are distorted by off-set counts ("dark currents") and pixel-dependent sensitivities. Using the corresponding recorded dark-field image D and flat-field image F, P can be corrected and converted into a beam intensity loss image I following the Beer-Lambert law as

$$I = -\log\left(\frac{P - D}{F - D}\right). \tag{1}$$

For each sample and each of the three source positions, a reconstruction was computed using the FDK algorithm[15] implemented in the ASTRA toolbox[16]. Then, the data from all source positions was combined to compute a high-quality reconstruction. This was done by solving a non-negativity constrained least-squares problem using 50 iterations of accelerated gradient descent[17]. The corresponding forward and backward projection operators were implemented using the CUDA kernels in the ASTRA toolbox. In both cases, we chose a volume of 501^3 voxels of size $100 \, \mu\text{m}^3$. The computation for one FDK reconstruction with the data from one orbit took about 24 s on an NVIDIA GeForce GTX 1070, the iterative reconstruction of the complete data 56 min. An example of the reconstructed volumes is shown in Fig. 5: In the FDK reconstructions from single orbit data, the image regions with low beam incident angles are reconstructed well while strong artefacts can be seen in regions with high beam angle. They are caused by a combination of two factors: First, the circular orbit associated with a cone shaped beam does not fulfill Tuy's condition[18], resulting in missing data in the measurement domain located around the rotation axis. Second, the FDK algorithm approximates the incoming beam by a collection of tilted fan-beams for each row of the flat detector. In contrast, the iterative reconstruction from the combined data is both sharp and artefact-free and can therefore be regarded as a ground truth reconstruction.

Data Records

The complete projection data for a single walnut and the corresponding reference reconstructions are shared as a single ZIP archive (ca. 6 GB per file). The 42 resulting ZIP files (named `Walnut1.zip` – `Walnut42.zip`, ca. 254.2 GB in total), were uploaded on zenodo (https://zenodo.org), and had to be split up into several bundles to with separate DOIs: Samples 1–8[19], samples 9–16[20], samples 17–24[21], samples 25–32[22], samples 33–37[23] and samples 38–42[24]. Note, however, that each ZIP file can be downloaded separately via zenodo's web interface.

The ZIP file for the ith sample, `Walnut<i>.zip`, contains a folder `Walnut<i>/` with the sub-folders `Projections/` and `Reconstructions/`:

- `Projections/tubeV<j>/` contains the measured projection data with the source at position `j`, where `j = 1/2/3` corresponds to the high/middle/low source position (cf. Fig. 3). Each of these folders contains the files:
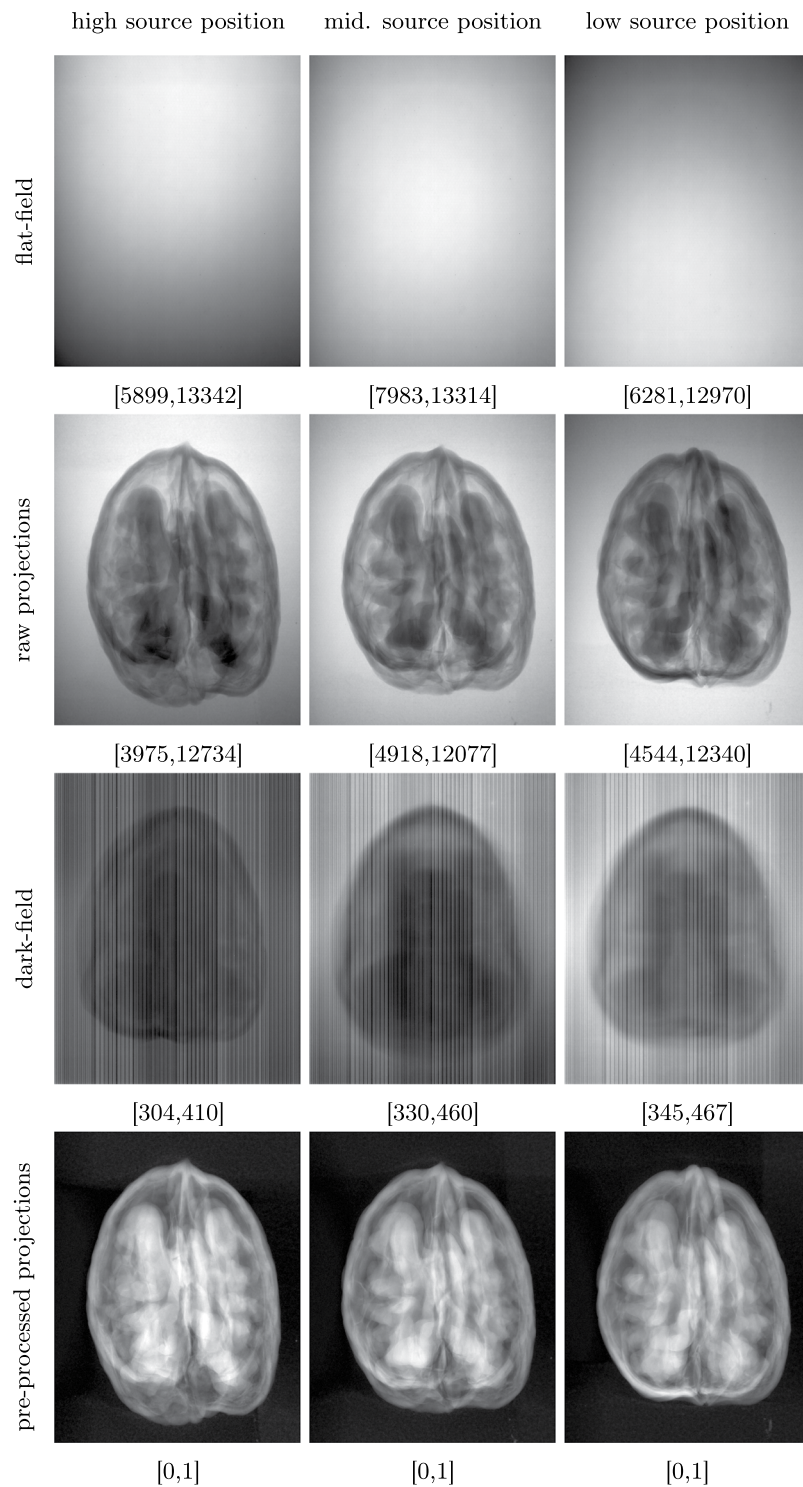
Fig. 4 Examples of the collected projections. From left to right, the position of the source varies. The dynamic range is indicated below each image.

- `di000000.tif` is a TIFF file containing the dark-field measurement (cf. Fig. 4).
- `io000000.tif` and `io000001.tif` are TIFF files containing the flat-field measurements before and after the orbit was scanned (cf. Fig. 4).
- `scan_<k>.tif` is a TIFF file containing the projection measurement at angle k (cf. Fig. 4).
- `scan_geom_original.geom` and `scan_geom_corrected.geom` are text files describing the acquisition geometry of each angular projection. Their format and usage are explained in more detail in the following sections.
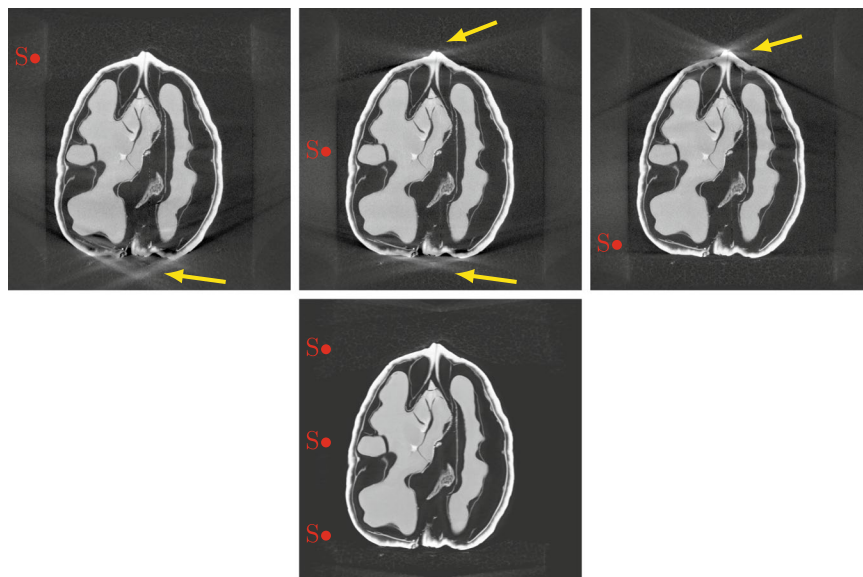
Fig. 5 Vertical slice through reconstructed volumes from a single sample. Red dots indicate the source height for the circular orbit used for the reconstruction. Top row: FDK reconstruction from top, middle, and low source position. Yellow arrows point at the high cone angle artefacts. Bottom row: Iterative reconstruction from combined measurements.

- `data settings XRE.txt` and `scan settings.txt` are text files automatically generated by the FleX-ray scanning software containing scan settings such as motor positions, source power or camera exposure time. We included them for completeness.
- `script_executed.txt` is a text file automatically generated by the FleX-ray scanning software containing a copy of the script executed by the scanner. We included it for completeness.

- `Reconstructions` contains the reference reconstruction as described above, stored as TIFF files each containing a single x-slice of the volume:

  - `fdk_pos<j>_<k>` contains the kth x-slice of the FDK reconstruction computed from the projection data acquired at source position j (cf. Fig. 5).
  - `full_AGD_50_<k>` contains the kth x-slice of the ground truth reconstruction computed by 50 iterations of accelerated gradient descent (cf. Fig. 5).

Technical Validation

The FleX-ray scanner is subject to regular maintenance and calibration. Furthermore, a visual inspection of all projections for each sample was carried out to ensure that the collected data does not suffer from over-saturation and that the sample was always in the field of view. The reconstructed volumes were inspected to ensure that the correction of geometric distortions, such as in-plane rotation tilt of the detector, was successful. For the iterative reconstruction from the combined data (ground truth reconstruction), the registration of the scanning geometries from the single orbits had to be corrected manually due to mechanical inaccuracies in the motors positions reported by the scanner. For this, three volumes corresponding to the three orbits were reconstructed first and then manually co-registered using rigid transformations. Corresponding corrected geometry description text files that are used in the combined reconstruction are provided (`scan_geom_corrected.geom`). Samples for which this procedure did not succeed were discarded. For completeness, the original geometry description text files as deduced from the reported motor positions are also provided (`scan_geom_original.geom`).

Usage Notes

Projection data. The projection data for each sample are shared as a collection of 16 bit unsigned integer TIFF files containing the raw photon counts per detector pixel. They can be interpreted and manipulated by most common image visualization software such as ImageJ[25] or scientific computing languages such as MATLAB or Python, e.g., through the matplotlib module for the latter. In order to be used by most tomographic reconstruction algorithms, they need to be pre-processed as described above and exemplified in the provided scripts. Each row of the geometry description files (`scan_geom_*.geom`) describes the geometry of one of the acquired projections by 12 floating point numbers: source x position, source y position, source z position, detector center x position, detector center y position, detector center z position, detector 3D basis vector from pixel $(0, 0)$ to pixel $(1, 0)$, and detector 3D basis vector from pixel $(0, 0)$ to pixel $(0, 1)$. This parametrization is illustrated in Fig. 6.
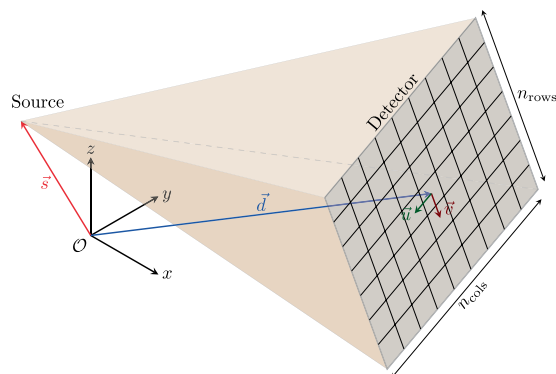
Fig. 6 Parametrization of the cone-beam geometry: Each projection is described by $(s_x, s_y, s_z, d_x, d_y, d_z, u_x, u_y, u_z, v_x, v_y, v_z)$.

Reconstructed volumes. In principle, the four reconstructions described in the previous sections (cf. Fig. 5) can be computed from the projection data with the scripts provided. Depending on the available computational resources this may, however, require a lot of computing time. For this reason, we share the reconstructions, too. They can also be used as a comparison ground to test novel reconstruction algorithms, or as an image collection for image analysis tasks. Each volume is released as a collection of 32 bit floating point TIFF files, where every single file is one axial slice through the volume as described above. As for the projection data, open source software is available for visualization and manipulation of such files.

Further usage. The reconstruction scripts can easily be modified to generate different kind of artefacts and tackle different problems related to tomographic reconstruction. To create a limited or sparse-angle (low-dose) problems, one can simply load subsets of the projection data. To mimic a super-resolution experiment, the projection data can be artificially binned into larger pixels. In every case, the iterative reconstruction from the full data set can be used as a ground truth.

Code availability

Python and MATLAB scripts for loading, pre-processing and reconstructing the projection data in the way described above are published on github: https://github.com/cicwi/WalnutReconstructionCodes.

They make use of the ASTRA toolbox, which is openly available on www.astra-toolbox.com or accessible as a conda package (use `conda install -c astra-toolbox/label/dev astra-toolbox` to install the development version). ASTRA is currently only fully supported for Windows and Linux. Installing it on Mac OS is possible but in the current state very involved and version-dependent. For obtaining a comparable scaling of the image intensities between FDK and iterative reconstructions, it is required to use a development version of the ASTRA toolbox more recent than 1.9.0 dev. For each dataset, a text file containing information about motor positions (source 3D position, detector position and detector orientation) is provided and used by the aforementioned Python/MATLAB scripts to set up the reconstruction geometry. All reference reconstructions provided have been computed with the Python scripts. Furthermore, while the scripts allow to sub-sample the projections and to choose a different image resolution, the reference reconstructions were computed with all projections and within a volume of 501^3 voxels of size $100\,\mu\text{m}^3$ as mentioned above.

References

1. Hsieh, J. A practical cone beam artifact correction algorithm. *IEEE Nuclear Science Symposium* **2**, 15/71–15/74 (2000).
2. Dennerlein, F., Noo, F., Schöndube, H., Lauritsch, G. & Hornegger, J. A factorization approach for cone-beam reconstruction on a circular short-scan. *IEEE Transactions on Medical Imaging* **27**(7), 887–896 (2008).
3. Zhang, Z. *et al.* Artifact reduction in short-scan CBCT by use of optimization-based reconstruction. *Physics in Medicine and Biology* **61**(9), 3387 (2016).
4. Koivisto, J., Eijnatten, M., Kiljunen, T., Shi, X. & Wolff, J. Effective Radiation Dose in the Wrist Resulting from a Radiographic Device, Two CBCT Devices and One MSCT Device: A Comparative Study. *Radiation Protection Dosimetry* **179**, 58–68 (2017).
5. Wang, G., Ye, J. C., Mueller, K. & Fessler, J. A. Image reconstruction is a new frontier of machine learning. *IEEE Transactions on Medical Imaging* **37**(6), 1289–1296 (2018).
6. Ravishankar, S., Ye, J. C. & Fessler, J. A. Image Reconstruction: From Sparsity to Data-adaptive Methods and Machine Learning. *Proceedings of the IEEE* (Early access), 1–24 (2019).
7. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE* **86**(11), 2278–2324 (1998).
8. Coban, S. B. SophiaBeads Datasets Project Documentation and Tutorials. *MIMS ePrints* **26**, 1–22 (2015).
9. Jørgensen, J. S., Coban, S. B., Lionheart, W. R. B., McDonald, S. A. & Withers, P. J. SparseBeads data: benchmarking sparsity-regularized computed tomography. *Measurement Science and Technology* **28**(12), 124005 (2017).
10. Singh, K. *et al.* Time-resolved synchrotron X-ray micro-tomography datasets of drainage and imbibition in carbonate rocks. *Scientific Data* **5**, 180265 (2018).
11. De Carlo, F. *et al.* Tomobank: a tomographic data repository for computational X-ray science. *Measurement Science and Technology* **29**(3), 034004 (2018).

12. Hämäläinen, K. *et al*. Tomographic X-ray data of a walnut. Preprint at, https://arxiv.org/abs/1502.04064 (2015).
13. McCollough, C. TU-FG-207A-04: Overview of the Low Dose CT Grand Challenge. *Medical Physics* **43**(6 Part 35), 3759–3760 (2016).
14. Pelt, D., Batenburg, K. J. & Sethian, J. Improving Tomographic Reconstruction from Limited Data Using Mixed-Scale Dense Convolutional Neural Networks. *Journal of Imaging* **4**(11), 128 (2018).
15. Feldkamp, L. A., Davis, L. C. & Kress, J. W. Practical cone-beam algorithm. *Journal of the Optical Society of America A* **1**(6), 612–619 (1984).
16. Van Aarle, W. *et al*. Fast and flexible X-ray tomography using the ASTRA toolbox. *Optics Express* **24**(22), 25129–25147 (2016).
17. Chambolle, A. & Pock, T. An introduction to continuous optimization for imaging. *Acta Numerica* **25**, 161–319 (2016).
18. Tuy, H. K. An inversion formula for cone-beam reconstruction. *SIAM Journal of Applied Mathematics* **43**(3), 546–552 (1983).
19. Der Sarkissian, H. *et al*. Cone-Beam X-Ray CT Data Collection Designed for Machine Learning: Samples 1–8. *Zenodo*. https://doi.org/10.5281/zenodo.2686725 (2019).
20. Der Sarkissian, H. *et al*. Cone-Beam X-Ray CT Data Collection Designed for Machine Learning: Samples 9–16. *Zenodo*. https://doi.org/10.5281/zenodo.2686970 (2019).
21. Der Sarkissian, H. *et al*. Cone-Beam X-Ray CT Data Collection Designed for Machine Learning: Samples 17–24. *Zenodo*. https://doi.org/10.5281/zenodo.2687386 (2019).
22. Der Sarkissian, H. *et al*. Cone-Beam X-Ray CT Data Collection Designed for Machine Learning: Samples 25–32. *Zenodo*. https://doi.org/10.5281/zenodo.2687634 (2019).
23. Der Sarkissian, H. *et al*. Cone-Beam X-Ray CT Data Collection Designed for Machine Learning: Samples 33–37. *Zenodo*. https://doi.org/10.5281/zenodo.2687896 (2019).
24. Der Sarkissian, H. *et al*. Cone-Beam X-Ray CT Data Collection Designed for Machine Learning: Samples 38–42. *Zenodo*. https://doi.org/10.5281/zenodo.2688111 (2019).
25. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nature methods* **9**(7), 671–675 (2012).

Author contributions

H.D.S., F.L., M.v.E. and K.J.B. conceptualized the study and designed the experiment. H.D.S., G.C. and S.B.C. set up the experiment and performed the data acquisition. H.D.S. performed the data processing, inspection and geometry correction. H.D.S. and F.L. wrote the reconstruction scripts and the the main parts of the manuscript. All authors contributed to the discussion and finalization of the manuscript and approved it.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.D.S. or F.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.