U-Net Segmentation Methods for Variable-Contrast XCT Images of Methane-Bearing Sand Using Small Training Datasets

Authors

Fernando J. Alvarez-Borges[ab], Oliver N.F. King[b]*, Bangalore N. Madhusudhan[a], Thomas Connolley[b], Mark Basham[c] and Sharif Ahmed[b]

[a] University of Southampton, Southampton, SO17 1BJ, United Kingdom

[b] Diamond Light Source (United Kingdom), Didcot, OX11 0DE, United Kingdom

[c] The Rosalind Franklin Institute, OX11 0QS, UK

Correspondence email: olly.king@diamond.ac.uk

Funding information Natural Environment Research Council (grant No. NE/K00008X/1 to Bangalore N. Madhusudhan).

Synopsis U-Nets were used to perform multiphase segmentation of synchrotron XCT scans of CH_4-bearing sand. These networks were trained on very small targeted datasets but still out-performed mainstream thresholding and watershed methods. They could also produce accurate segmentations for completely new data without additional training.

Abstract Methane (CH_4) hydrate dissociation and CH_4 release are potential geohazards currently investigated using X-ray computed tomography (XCT) imaging in laboratory experiments. Image segmentation constitutes an important data processing step for this type of research, but it is often time consuming, computing resource-intensive and operator-dependent. Furthermore, segmentation procedures are frequently tailored for each XCT dataset due to differences in image characteristics, such as greyscale contrast variations. To address these issues, an investigation has been carried out using U-Nets, a class of Convolutional Neural Network, to segment synchrotron radiation XCT (SRXCT) images of CH_4-bearing sand during hydrate formation. Emphasis was given to CH_4 gas bubbles, due to their paucity and low contrast. Three U-Net deployments previously untried for this task were assessed: (1) a bespoke 3D hierarchical method, (2) a 2D multi-label, multi-axis method and (3) RootPainter, an application that combines a 2D U-Net with interactive corrections. U-Nets were trained using very small hand-annotated datasets to reduce operator time. Results show high segmentation accuracy and consistency, with RootPainter slightly outperforming the alternative approaches and all three methods

surpassing mainstream watershed and thresholding techniques. Greyscale contrast between material phases was found to affect segmentation performance, with the lowest metrics corresponding to data exhibiting the lowest contrast. Segmentation accuracy affected derived parameters such as CH_4-saturation and porosity, but errors were small compared with gravimetric methods. It was also found that U-Net models trained on low greyscale contrast images could be used to segment higher-contrast datasets and also data collected at a different facility, thereby demonstrating model portability. Such portability is anticipated to be advantageous when the segmentation of large XCT datasets needs to be delivered over short timespans.

Keywords: U-Net, sediment microstructure, microtomography, segmentation.

1. Introduction

Deep sea sediments and permafrost host large quantities of methane (CH_4), an energy source and potent greenhouse gas that may be a contributor to climate change (Dean *et al.*, 2018; IPCC, 2013). Much of this CH_4 is present as hydrates (clathrates), that is, solid crystalline lattices of water at low temperature and high pressures that enclose CH_4 molecules. 164 m^3 of CH_4 gas at normal temperature and pressure can be stored in one m^3 of hydrate (Kvenvolden, 1993). However, the extent of the world-wide CH_4 hydrate inventory is subject to considerable uncertainty (James *et al.*, 2016; Ruppel & Kessler, 2017). This is in part due to discrepancies between measurements produced by geophysical and electrical resistivity methods (Sahoo, Marín-Moreno, *et al.*, 2018; Yokohama *et al.*, 2011), which are potentially associated with hydrate and CH_4 gas distribution heterogeneity in the host soils (Sahoo, Madhusudhan, *et al.*, 2018). Uncertainties regarding the global CH_4 hydrate inventory affect resource estimation and CH_4 emission prediction models (Moridis *et al.*, 2011; Ruppel & Kessler, 2017; Saunois *et al.*, 2020). CH_4 hydrate formation and dissociation has also been associated with changes in the mechanical characteristics of the host sediment. For instance, hydrates may strengthen and stiffen the sediment matrix by creating inter-grain cementation bonds (Madhusudhan *et al.*, 2019; Song *et al.*, 2019). This is speculated to lead to, for example, underwater slides that may trigger tsunami or damage seabed infrastructure such as cables and pipelines (Maslin *et al.*, 2010; Mienert, 2009; Vanneste *et al.*, 2014).

Recently, researchers have shown that X-ray computed tomography (XCT) can be used to successfully detect hydrate and CH_4 gas bubble distribution heterogeneity and characterise changes in sediment microstructure associated with hydrate formation and dissociation (Holland & Schultheiss, 2014; Kerkar *et al.*, 2014; Lei *et al.*, 2018; Sahoo, Madhusudhan, *et al.*, 2018). This has been possible in great part due to advancements in image segmentation techniques. Segmentation is the process of classifying 2D pixels or 3D voxels into regions, for example, the solids (e.g., soil grains and cement bonds), liquids (e.g., water or brine) and gases (e.g., air or CH_4) present in an XCT image of a

geomaterial sample. Microstructural parameters such as porosity and grain and pore size, shape and orientation can then be derived from the segmented image, as well as volumetric quantities such as CH_4 gas and hydrate saturation ratios.

Some of the most common segmentation techniques used in geomechanics and geoscience are greyscale thresholding and watershed algorithms (Fonseca *et al.*, 2009; Iassonov *et al.*, 2009). The former involves the selection of a greyscale range to classify pixels or voxels into regions of interest. Watershed algorithms redefine the image as a geographical map, where greyscale intensities form topographical elevations and catchment basins. Pixel/voxel markers within these basins are used to define the materials or 'labels' present in the image, and the algorithm then morphologically dilates these markers until they 'fill' their catchment basins (Rogowska, 2000; Zhang *et al.*, 2014). Greyscale range determination in the case of thresholding techniques and marker grey value and location in the case of watershed techniques are operator and/or method dependent (Baveye *et al.*, 2010; Fonseca *et al.*, 2009; Koyuncu *et al.*, 2012). The values assigned to these parameters also depend on the recorded greyscale contrast, which is highly reliant on the X-ray imaging instrument used and how it was optimised (Brunke *et al.*, 2008). Sample heterogeneity or density changes during an in-situ experiment will further introduce contrast variability in space and time (Fonseca *et al.*, 2009; Kong & Fonseca, 2018). As a result, thresholding and watershed segmentation are typically optimised per XCT scan and objective comparison is difficult given that the data treatment varies between datasets. These issues often result in segmentation procedures in geomechanics and geoscience that are highly demanding of computing resources and operator time.

Novel alternative approaches have employed machine learning to segment the multiple material phases present in XCT images of soil and rock samples (Chauhan, Rühaak, Anbergen, *et al.*, 2016; Chauhan, Rühaak, Khan, *et al.*, 2016). For these applications, segmentations are produced via a mathematical model optimised or 'trained' using a series of 'ground truth' example segmentations of XCT images provided by the user. Within the realm of machine learning, convolutional neural networks (CNNs) are a class of deep neural networks that employ multiple convolutional layers where the filters ('kernels') used to separate image features are learned (Krizhevsky *et al.*, 2017). Researchers have recently begun exploring the application of CNNs to segment XCT images of soil and rock (Douarre *et al.*, 2018; Karimpouli & Tahmasebi, 2019; Phan *et al.*, 2021; Varfolomeev *et al.*, 2019).

U-Nets are a class of CNN originally designed to segment biomedical images (Ronneberger *et al.*, 2015). The U-Net architecture is composed of downsampling (encoding/contracting) and upsampling (decoding/expanding) paths. The former reduces the spatial dimensions of the data while increasing feature information while the latter recombines spatial and feature data to generate the label image. The encoding and decoding sections of the network are linked by connections that can feed the output from the contracting path directly into the corresponding level of the expanding path. This allows the

transfer of spatial information and the preservation of fine-grained details in the output label image. A limitation to the implementation of U-Nets (and CNNs in general) to segment XCT images of soil and rock is the preparation of training and validation datasets, which often require labour-intensive manual segmentation (hand annotation) of many images.

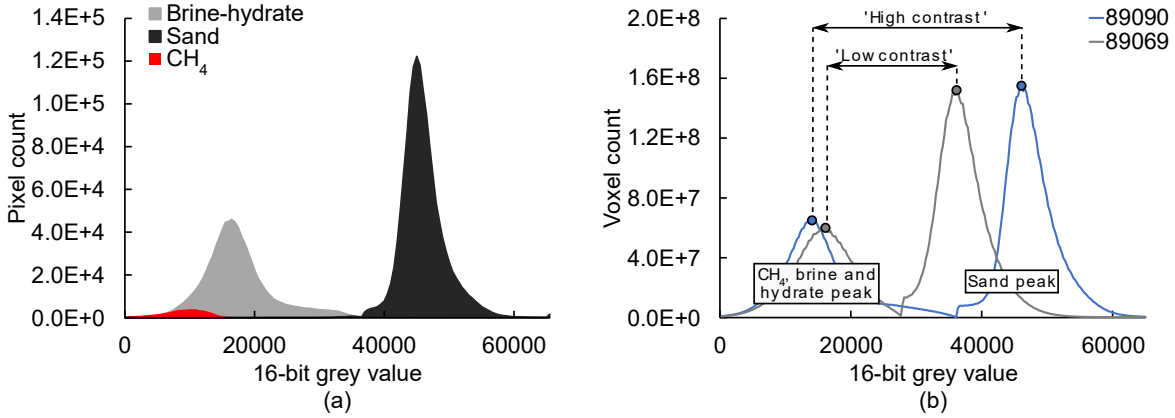

Figure 1 Grey value histograms of reconstructed and post-processed SRXCT images: (a) of XY slice 1050 of scan 89062, showing the frequency distribution of pixels for each material; (b) of two whole 3D images showing the grey value difference between histogram peaks as a measure of image contrast.

This paper examines the use of U-Nets trained on very small sets of ground truth data to segment a large number of synchrotron radiation XCT (SRXCT) images of CH_4-bearing sand. The suitability of such time- and resource-saving approach has not been examined previously. Three different U-Net implementation strategies, also previously untried in the field of geomechanics and geoscience research, have been applied. Two of these strategies are entirely novel. The U-Net segmentation procedures targeted the three main material phases present in the images: sand, CH_4 gas bubbles and brine combined with hydrates. Special focus has been given to the CH_4 gas phase, as it not only exhibited low contrast with regards to the brine-hydrate phase but was also uncommon in the data compared to the other materials, as shown in Figure 1(a). Hydrates were not targeted separately as the principal aim of the experiment was to study bubble morphology without the presence of contrast agents foreign to the CH_4-sea water-soil model such as KI and Xe as done in previous work (Lei *et al.*, 2018; Sell *et al.*, 2016). The SRXCT data was obtained from in situ imaging of hydrate formation and dissociation experiments. The reconstructed volumes exhibited different greyscale contrast amongst them. Furthermore, contrast between the CH_4 gas and pore water phases present in the images was low, as shown in the image histogram for a reconstructed slice in Figure 1(a). This rendered the use of conventional thresholding or watershed techniques largely unsuitable. The aim of this investigation was thus to determine if U-Nets can accurately segment XCT images of soil samples with varying greyscale contrast between material phases using only a small number of training and validation images, therefore reducing operator and computing time and allowing for objective data

comparison. The starting hypotheses were (1) that U-Net models trained on a small portion of the reconstructed SRXCT 3D image can be used to accurately segment the entire volume, (2) that segmentation accuracy is directly linked to greyscale contrast between materials, and (3) that accurate U-Net segmentation models produced from training on a given SRXCT dataset can deliver accurate segmentations for similar datasets without additional training (model portability).

2. Materials and methods

2.1. Methane gas hydrate formation and dissociation experiments

A custom rig designed and manufactured by Sahoo, Madhusudhan, *et al.* (2018) for in situ SRXCT imaging of gas hydrate formation and dissociation was used in the present study. The rig is made of polyether ether ketone (PEEK) and consists of a monolithic 2 mm internal diameter by 23 mm tall cylindrical vessel with 0.8 mm thick walls and an enlarged base, as shown in Figure 2. The soil sample is placed through the bottom of the rig. The pore fluid injection pipe is connected to this inlet, as depicted in Figure 2. The rig features thermocouples at the base of the scan zone shown in Figure 2 to measure sample temperature. The SRXCT imaging zone in this study corresponds to a vertically centred 1.755 mm-tall region within the 10 mm-tall scan zone.

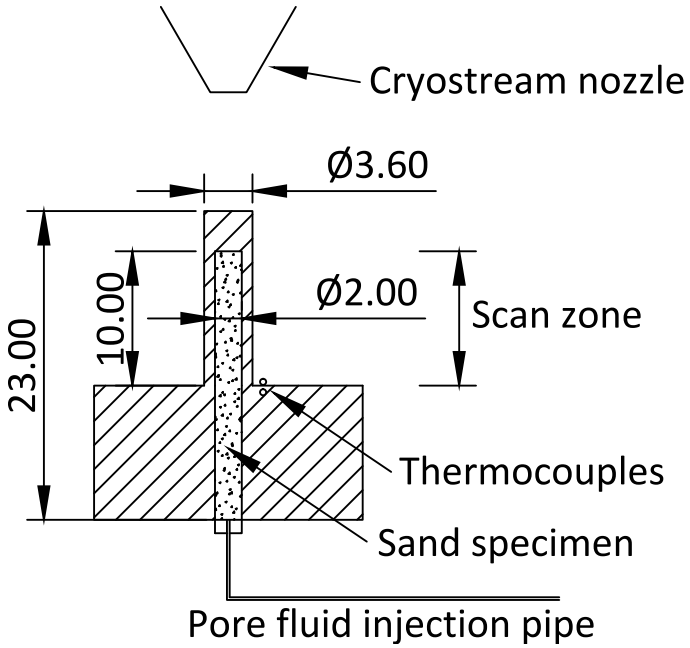

Figure 2 Cross-section sketch of hydrate test rig. Monolithic PEEK element denoted by hatched area. All units mm.

Leighton Buzzard sand Fraction E (LBE) with mean grain diameter of 100 µm was used as surrogate marine sediment. LBE is an angular silica sand widely used as a standard laboratory material in geomechanics research. The sand was tamped into the PEEK vessel to a target porosity of 35%. A

vacuum pressure of less than 1 Pa was applied through the injection pipe to reduce air presence in the pore space. A calculated volume of brine solution (3.5% NaCl by weight, representative of deep ocean water; Brown (2016)) was thereafter injected into the sample, such that approximately 90% of the pore volume became saturated. CH_4 gas was then injected at 10 MPa and the valve to the sample closed. The sample was gradually cooled to a target constant temperature of 2 °C using a N2 cryostream. This thermobaric condition enabled hydrate formation in the pore space instead of ice. The target temperature was maintained for 30 hours to complete the hydrate formation process (Madhusudhan *et al.*, 2019).

2.2. Synchrotron X-ray computed tomography

2.2.1. Set-up and image acquisition

Data was collected on beamline I13-2 at Diamond Light Source (DLS). Scans were performed using a polychromatic 'pink beam' at 30 keV peak energy. The detector system used was a scintillator-coupled pco.edge 5.5 camera fitted with a 4x optic magnification lens, resulting in an effective pixel size of 0.8125 µm. The X-ray projection size was 2560×2160 pixels (width × height).

Scans were carried out in-situ at various time intervals after reaching 2 °C. The number of projections and the exposure time per projection varied amongst scans to reduce acquisition times at specific moments of the CH_4 hydrate formation process. Table 1 correlates each scan discussed in this paper with the time after the start of the 30-hour sustained 2 °C period, as well as the scan specifications used.

Table 1 SRXCT scan summary.

Dataset	Time at 2°C (h)	Number of projections	Exposure time per projection (ms)
89062	0.00	1501	200
89064	1.53	1501	200
89069	5.38	3001	30
89075	10.72	3001	30
89090	20.77	1501	30
89113	30.02	1501	30

2.2.2. Tomographic reconstruction and post-processing

Tomographic reconstruction was carried out using Savu software (Wadeson *et al.*, 2019; Atwood *et al.*, 2015; Wadeson & Basham, 2016). Two Savu reconstruction pipelines were used: one with and one without Paganin phase enhancement (Paganin *et al.*, 2002). These pipelines were labelled 'phase contrast' (Figure 3(b)) and 'absorption contrast' (Figure 3(a)), respectively. Both pipelines implemented filtered back-projection reconstruction (Ramachandran & Lakshminarayanan, 1971; van Aarle *et al.*, 2016) and pre-reconstruction algorithms for speckle and ring artefact suppression (Atwood *et al.*, 2015; Titarenko *et al.*, 2010) and the automatic determination of the centre of rotation (Vo *et al.*, 2014). Further processing was carried out on the output from both reconstruction pipelines using Fiji (Schindelin *et al.*, 2012; Schneider *et al.*, 2012). This consisted in:

1. The application of a median filter of kernel size 3 to the absorption volume and the halving of the resulting greyscale values.
2. The application of an unsharp mask filter of radius 3 and weight 0.70 to the phase contrast volume.
3. The elementwise averaging of both volumes.

This procedure resulted in a single reconstructed volume with clear edge detail and phase contrast (Figure 3(c)).

Finally, to mitigate the halo-like or 'cupping' artefact caused by the preferential attenuation of lower-energy X-rays close to the specimen surface, known as beam hardening, as well as by truncation artefacts introduced by attenuation from sample regions outside the field of view (Hsieh, 2015; Kalender, 2011),), each slice was convolved with two mollifier functions with an inverse shape to that of the cupping artefact. This flattened the horizontal (XY) grey value profile of each slice. A circular mask with a radius of 1100 pixels was then applied to remove voxels at the outer edges of the field of view (FOV), which were resistant to cupping correction. An example output slice is presented in Figure 3(d).

As outlined in Section 1, limited greyscale contrast between the CH_4 gas and the brine-hydrate phase persisted after reconstruction and post-processing. Distinction between these two phases became increasingly difficult as the distance between the 3D image histogram peaks for the sand and non-sand phases reduced, as exemplified in Figure 1(b). This distance is therefore used in this paper as an overall measure for image contrast, with regards to the ease with which the material phases could be identified and segmented. Considering this, 'intermediate contrast' dataset 89062 was selected initially to investigate the suitability of U-Nets to perform segmentations.

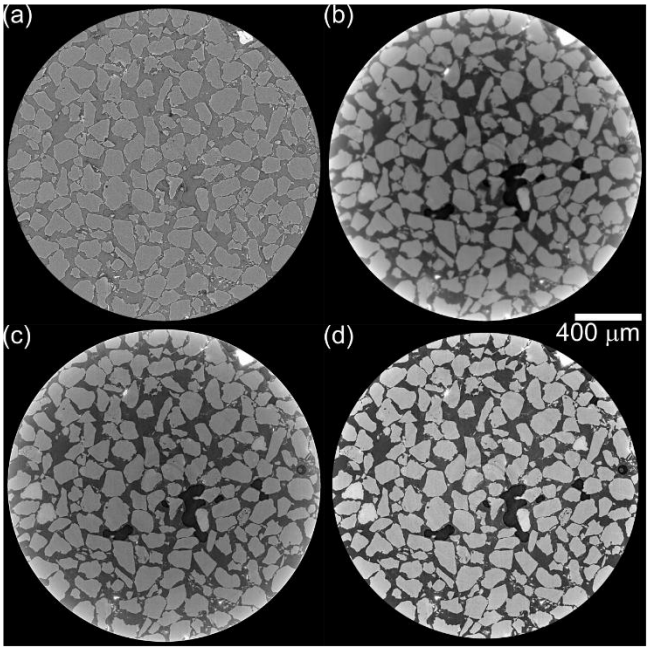

Figure 3 Slice 1050 of dataset 89062 showing the output of the reconstruction and post-processing stages: (a) reconstruction through absorption contrast pipeline; (b) reconstruction through phase contrast pipeline; (c) Output from filtering and volume averaging; (d) Cupping correction output.

2.3. U-Net Segmentation

Three different methodologies were used to create trained U-Net models to segment the three main material phases present in the images: sand, brine-hydrates and CH_4 gas. These were:

1. A 3D hierarchical approach where two separate 3D U-Net models were trained to perform binary segmentations: On the sand phase vs the others and the CH_4 gas phase vs the others.
2. A 2D multi-label and multi-axis approach where a single 2D U-Net was trained to classify the three labels. The encoder section of this U-Net implementation was pre-trained on the ImageNet dataset (Russakovsky *et al.*, 2015), meaning that the network should only require a small amount of 'transfer' training in order to achieve acceptable results on new data.
3. RootPainter software, which uses a graphical user interface (GUI) and human intervention by interactive corrections to train a lightweight binary 2D U-Net model.

The U-Net models produced by each method were used to segment a 1554×1554×2000 voxel region of the 2560×2560×2000 reconstructed and post-processed volumes. This region was inscribed within the cylindrical FOV of the post-processed volumes and omitted the black pseudo-background generated during reconstruction. Figure 4(a) shows the 1554×1554×2000 volume for dataset 89062. All 1554×1554×2000 images discussed in this paper are available in (Alvarez-Borges *et al.*, 2021). It

is emphasised that the segmentation of the sand phase via U-Nets was done to assess the multi-label segmentation capacity of the algorithm. Due to its uniformly high-contrast and well-defined edges, sand could be easily segmented with any 'standard' method, for example, thresholding.

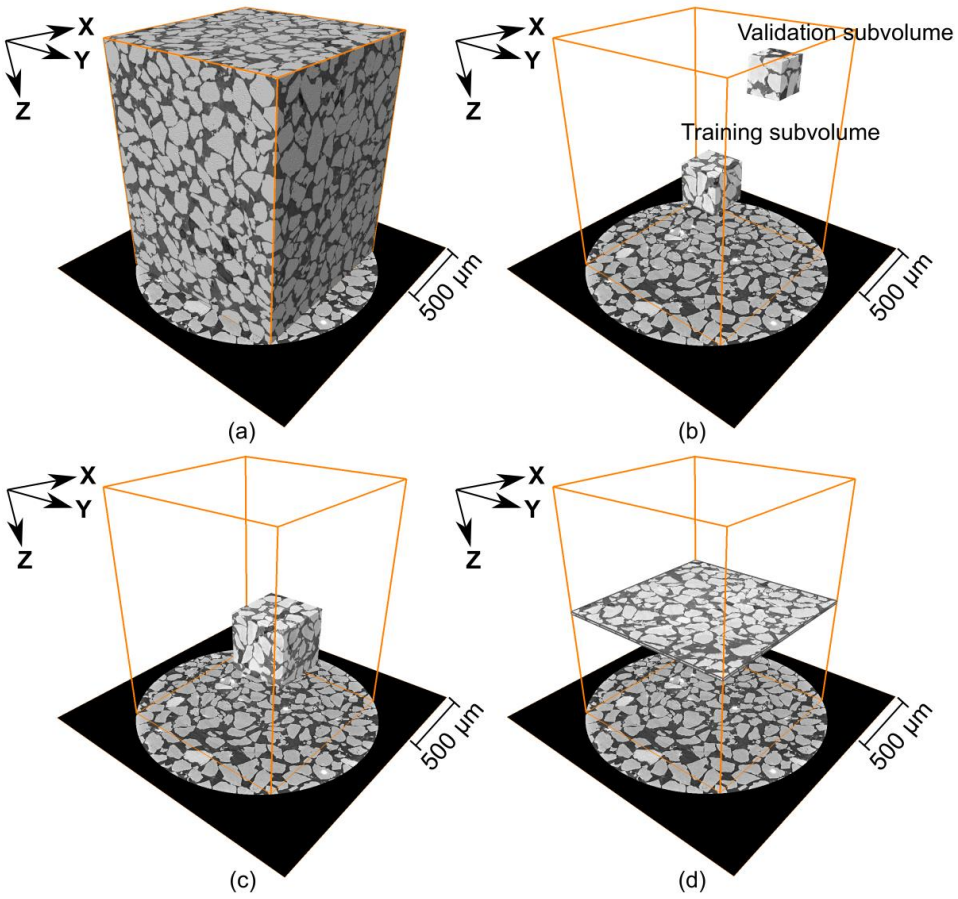

Figure 4 For dataset 89062: (a) 1554×1554×2000 voxel central region used for U-Net segmentation; (b) Location of 384×384×384 and 256×256×256 voxel training and validation volumes, respectively; (c) 572×572×572 voxel training and validation subvolume; (d) Central 40 slices used for quantitative analysis.

2.3.1. Training and validation data

The U-Net training procedures required both greyscale and label datasets. The latter was the 'ground truth' information used during training and validation. Label data was produced by hand-annotating the sand, CH_4 gas and brine-hydrate in the greyscale data using Avizo Lite® software. This was carried out on small subregions of the 1554×1554×2000 volumes to reduce labelling time. The 3D hierarchical approach used a 384×384×384 voxel (hereafter referred to as $[384]^3$) training sub-volume and a 256×256×256 voxel (hereafter referred to as $[256]^3$) validation sub-volume, selected from two different regions of the 3D image (Figure 4(b)). RootPainter requires 2D label images (slices) of at least 572×572 pixels in size for both training and validation, as explained later in Section 2.3.4. Therefore, a 572×572×572 voxel (hereafter referred to as $[572]^3$) sub-volume was delimited for this

purpose (Figure 4(c)). The same $[572]^3$ sub-volume was used to train the 2D multi-label models, while the $[256]^3$ sub-volume was used for validation. It should be noted that both the 3D hierarchical and 2D multi-label methods are able to use smaller training volumes, for example 128×128×128 voxels (hereafter referred to as $[128]^3$). However, it was deemed that such a small dataset size would result in U-Net models that were vastly overfitted to that particular region of the 3D image, weakening the model's ability to generalise to new data.

The training and validation sub-volume coordinate origins relative to the global origin of the reconstructed 2560×2560×2000 dataset are listed in Table 2. The global coordinate system origin is indicated in Figure 4, which also presents the location of the training and validation volumes (Figure 4(b-c)). All training, validation and segmented data used in this investigation are available in (Alvarez-Borges *et al.*, 2021).

Table 2 Training and validation sub-volume origin voxel coordinates relative to global origin of the 2560×2560×2000 volume (shown in Figure 4).

Size (voxels)	X	Y	Z
256×256×256	1133	1753	50
384×384×384	1343	943	1158
572×572×572	1343	943	1158

2.3.2. 3D Hierarchical Segmentation

The 3D hierarchical U-Net model used was implemented in the Python library PyTorch (Paszke *et al.*, 2019) and based upon an existing implementation of a residual 3D U-Net from the literature (Lee *et al.*, 2017; Wolny *et al.*, 2020). The voxel datatype of the training and validation greyscale sub-volumes was rescaled from 16-bit to 8-bit depth, truncating values beyond 2.575 standard deviations of the mean to mitigate the skewing effect of outliers. The ground truth label volumes (with three labels: sand, brine-hydrates and CH_4 gas) were used to create separate binary label volumes, one with sand vs background and the other with CH_4 gas vs background. These volumes were used as the label data for training the separate binary 3D U-Net models.

Unlike the multilabel 2D U-Net implementation described later, this model had not been pre-trained on ImageNet and was therefore likely to require a larger amount of training data to reach a high segmentation accuracy. To overcome this, the TorchIO library (Pérez-García *et al.*, 2020) was used to sample$[128]^3$ sub-volumes from the $[384]^3$ greyscale training data and generate 48 sub-volumes with random noise, flips, blurs, affine, and elastic transformations to be used as an extended training data set for each training epoch (i.e., a full training cycle). In addition, the validation volume was randomly sampled, creating 12 sub-volumes for model validation after each training epoch.

During training, U-Net model parameter optimisation (i.e., the process of updating the model parameters on each training iteration) was carried out with a method known as AdamW (Loshchilov & Hutter, 2019). The learning rate, a parameter that controls the step size of the updates made by the optimiser, was cycled up and down every epoch to reduce the need to tune this parameter and to accelerate the training process (Smith, 2017). Binary cross entropy (BCE), a measure of the uncertainty between two data distributions, was used as the loss function (the function minimised by the optimiser during training). Training progress was monitored using Intersection Over Union (IOU) on the validation set as the evaluation metric. If either no improvement in validation loss occurred after 40 passes of the entire training dataset (epochs) or 100 epochs were completed, the model with the lowest validation loss was saved. This was aimed at preventing overfitting. Software source code for this method is available from King and Alvarez-Borges (2021).

When predicting segmentation for the 1554×1554×2000 greyscale volumes, two binary predictions were produced for each data set, one for sand vs background and the other for CH_4 gas vs background. These two label volumes were then combined using a label hierarchy: first, a new 1554×1554×2000 volume was created with all voxel labels set to brine-hydrates, then the labels corresponding to CH_4 gas were transferred from the CH_4 vs background prediction, and lastly the labels corresponding to sand were transferred from the sand vs background prediction.

2.3.3. 2D Multi-label segmentation

Training of the 2D U-Net with multiple labels was performed on the $[572]^3$ sub-volume using two approaches. The first mimicked that of RootPainter, described later, with the network being trained on horizontal 2D (XY) slices through the image volume. The second, multi-axis approach, utilised slices taken in the XY, XZ and YZ planes (coordinate system shown in Figure 4). Prior to training, the voxel intensities in the selected volume were rescaled to 8-bit depth, as done for the 3D hierarchical method. A 2D U-Net was used with a ResNet34 encoder (He *et al.*, 2016). This encoder was loaded with pre-trained weights from ImageNet. The model was created with Fastai (Howard & Gugger, 2020), a Python library which has a high-level interface that utilizes PyTorch. During training, default Fastai image transformations and augmentations were used. The loss function used was cross entropy (CE) and the evaluation metric used was the number of correctly labelled voxels expressed as a percentage. Training was carried out for 15 epochs.

For the single-axis implementation, the XY training stack and corresponding label stack of 572 images, with dimensions 572×572, were split into training (80%) and validation (20%) sets. When predicting the segmentation for the 1554×1554×2000 greyscale volumes, data was fed into the network in the form of 2000 XY slices of size 1554×1554 pixels.

For the multi-axis approach, the $[572]^3$ training data and corresponding label sub-volume were sliced into 2D images in the XY, XZ and YZ planes, resulting in 1716 training image and label pairs. These

images were also split into a training (80%) and validation (20%) set. When predicting the segmentations for the 1554×1554×2000 greyscale volumes, an averaging approach for data produced from each plane was used as described by (Tun *et al.*, 2020), but with a modification to take the multiple labels into account. In short, this averaging approach consisted in slicing, segmenting, and rotating the volume across the XY 4-fold symmetry plane and then splitting and hierarchically recombining the 12 resulting segmentation volumes so that two label volumes were obtained, one containing labels for sand vs background and the other for CH_4 vs background. These two binary label volumes were then combined into a multi-label volume as done for the data output from the 3D hierarchical method (Section 2.3.2).

Software source code for this method is available from King and Alvarez-Borges (2021).

2.3.4. RootPainter Segmentation

RootPainter (Smith & Ørting, 2020) is a client-server application originally developed to segment plant root features from photographs of soil profiles (Smith, Han, *et al.*, 2020; Smith, Petersen, *et al.*, 2020). The client GUI is employed to annotate 2D images from a dataset, such as a tomography image stack of horizontal (XY) slices, as in the present case. The tomography slices and corresponding annotations are then read by the server and used to train the segmentation model using a U-Net variant implemented in PyTorch and described by Smith, Han, *et al.* (2020) and Smith, Petersen, *et al.* (2020). To execute the training routine, the software creates a validation dataset by randomly selecting one annotation image out of every five created. The accuracy of the model produced at the end of each training epoch is evaluated using the F-score parameter described by Smith, Petersen, *et al.* (2020). At the end of each training epoch, F-score values for the current and previous model are compared and the one with the highest value is saved. Training is stopped if 60 epochs are completed without F-score improvements.

RootPainter uses interactive corrections. These are created by annotating image slices overlaid with the segmentation labels produced by the best model currently available. These corrective annotation slices are added to the training and validation datasets so that the five to one ratio is maintained.

At present, RootPainter can only predict binary segmentations ('foreground' vs 'background'). Therefore, it was initially used to segment the CH_4 gas phase only. The $[572]^3$ label sub-volume was used for training and validation.

Sparse annotations have been shown to produce better results than dense/intensive annotations when interactively training U-Net models (Smith, Han, *et al.*, 2020; Gonda *et al.*, 2017). Thus, arbitrarily sparsely annotated images were produced by converting all CH_4 gas labels into foreground and enclosing them with background labels that included brine-hydrate and sand pixels, as shown in Figure 5(a-b). This was done by morphologically dilating the CH_4 label of each slice in the training

dataset and re-labelling the added pixels as background. The annotated slices were then copied into annotation and validation directories, maintaining the five-to-one ratio. Training was initiated after copying the first batch of five images. Further batches were added if a training epoch finished without further improvements in F-score and the model could not segment the majority of CH_4 pixels, or if the erroneously segmented pixels were patently greater than the number of correctly segmented pixels, as shown in Figure 5(c). Corrective annotation was started after a training epoch had produced a model that segmented most of the CH_4 regions with a roughly equivalent number of erroneously labelled pixels, as presented in Figure 5(d-e). Once a model was produced that could segment CH_4 without evident erroneously labelled pixels, the software was left to carry on training until the 60-epoch limit was reached. The resulting model was then used to segment the 1554×1554×2000 SRXCT volume slice by slice.

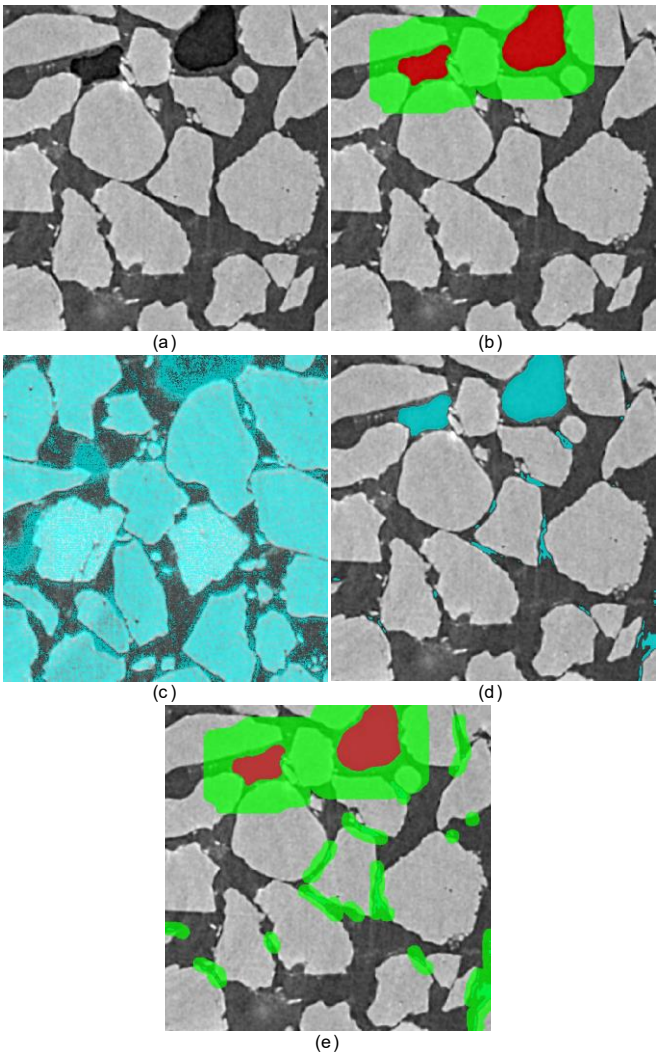

Figure 5 RootPainter usage example (on data from 89062): (a) XY slice from [572]3 sub-volume; (b) Slice annotations used for training and validation with CH4 (foreground) shown in red and background shown in green; (c) Initial segmentation output (blue) with a large number of erroneously

labelled voxels; (d) Improved segmentation with a small number of erroneously labelled voxels; (e) Annotative corrections on mislabelled voxels.

2.4. Thresholding and watershed segmentation

To compare the performance of the U-Net methods with conventional segmentation routines, the SRXCT data was segmented using manual and automatic thresholding, and the watershed method. Images were downsampled to 8-bit as in the U-Net methods described previously. A bilateral filter was used before segmentation to improve thresholding performance and mitigate over-segmentation (filter parameters were:100-pixel spatial kernel, 50-pixel window size, and a grey-value kernel of 30 counts; implemented in Python using the open-cv library, Bradski (2000); see e.g. Paris *et al.* (2009) for filter description).

Manual thresholding was carried out by selecting a single threshold value for all slices by visual inspection. Automatic thresholding was performed on a slice-by-slice basis using the multi-level Otsu method (Otsu, 1979) implemented using the scikit-image Python library (van der Walt *et al.*, 2014).

Watershed segmentation was carried out in Fiji using the morphological segmentation tool in the Morpholibj library (Legland *et al.*, 2016). It consisted in the application of a morphological gradient with radius of 1 and the automatic determination of markers by finding local minima (with a tolerance of 8 greyscale intensity values), prior to the watershed 'inundation' phase. The output label image contained different labels for all features in the greyscale input image, including sand and brine-hydrate. Labels corresponding to regions in the 8-bit volume with mean greyscale intensity values below 50 to 70, depending on the dataset, and above 130 were classified as CH_4 gas and sand, respectively. The remining voxels were classified as brine-hydrates.

2.5. Quantitative Analysis

The central 40 XY slices of the segmented 1554×1554×2000 volumes were compared with hand-annotated counterparts created in Avizo Lite® and considered to represent 'ground truth' labels. These slices do not intersect any of the training or validation subvolumes. These ground truth volumes are available in Alvarez-Borges *et al.* (2021). The previously mentioned IOU metric was used to evaluate segmentation performance. IOU is defined as:

$$\mathrm{IOU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \tag{1}$$

where TP refers to the number of voxels or pixels correctly predicted to correspond to the label of interest ('true positive'), and FP and FN are the number of voxels or pixels incorrectly predicted to be part of the label of interest ('false positive') and voxels/pixels incorrectly predicted to belong to any of the other material phases ('false negative'), in each case. A comparable analysis of U-Net accuracy has been done by, e.g., Karabağ *et al.* (2020) and Phan *et al.* (2021).

IOU returns a value between 0 and 1, where the latter corresponds to the scenario were the segmentation matches the validation image pixel by pixel (or voxel by voxel). In the following sections, quantitative analyses were carried out on a slice-by-slice basis (i.e., using pixel counts as input).

3. Results and Discussion

3.1. Segmentation Performance Comparison

Figure 6 compares the original and segmented central slice for the intermediate-contrast 89062 dataset, produced using the three U-Net methods (Section 2.3) and three standard methods (Section 2.4). Training and validation in both the 2D multi-label approach and RootPainter was carried out using XY slices only (i.e., single plane). Figure 7(a) presents accuracy metrics for the segmentation of CH_4 gas in the central 40 XY slices this dataset. It may be noted that RootPainter delivered slightly higher metrics than the other two U-Net methods, but this difference in performance cannot be readily identified in Figure 6. Figure 6 and Figure 7a also show that, for this dataset, watershed and manual thresholding methods return lower accuracy results than the U-Net approaches, and that Otsu-thresholding performed poorly. In fact, the Otsu approach consistently segmented the brine-hydrate and CH_4 gas as a single label, as evident in Figure 6e. This is chiefly due to the absence of well-defined inter-class variance extrema between these materials and the small relative size of CH_4 bubbles (Kittler & Illingworth, 1985; Lee *et al.*, 1990). In later comparisons, results from the Otsu method are omitted for this reason.

The slightly lower performance metrics observed in Figure 7(a) for the 3D hierarchical output, compared to that of RootPainter, may be attributed to the smaller training sub-volume used ($[384]^3$). To present a more balanced comparison, a further 3D hierarchical model was trained on a sub-volume of the same size as the one used for both 2D methods, i.e. $[572]^3$. This comparison is presented in Figure 7(b), where it is evident that RootPainter still outperformed the 3D hierarchical approach, though the difference between methods reduced.

Figure 7(a) and Figure 7(b) show that pre-training on the ImageNet database for the 2D multi-label method did not result in a significant segmentation performance advantage over the 3D hierarchical method. A similar outcome on the effect of transfer learning has been reported by He *et al.* (2019). They remarked that, ultimately, pre-training primes the U-Net for feature identification, which leads to fewer training iterations rather than greater segmentation accuracy. Such appears to be the present case, as the 2D multi-label approach produced similar results to the 3D hierarchical method with up to six times fewer training epochs, as shown in Table 3 and Table 4.

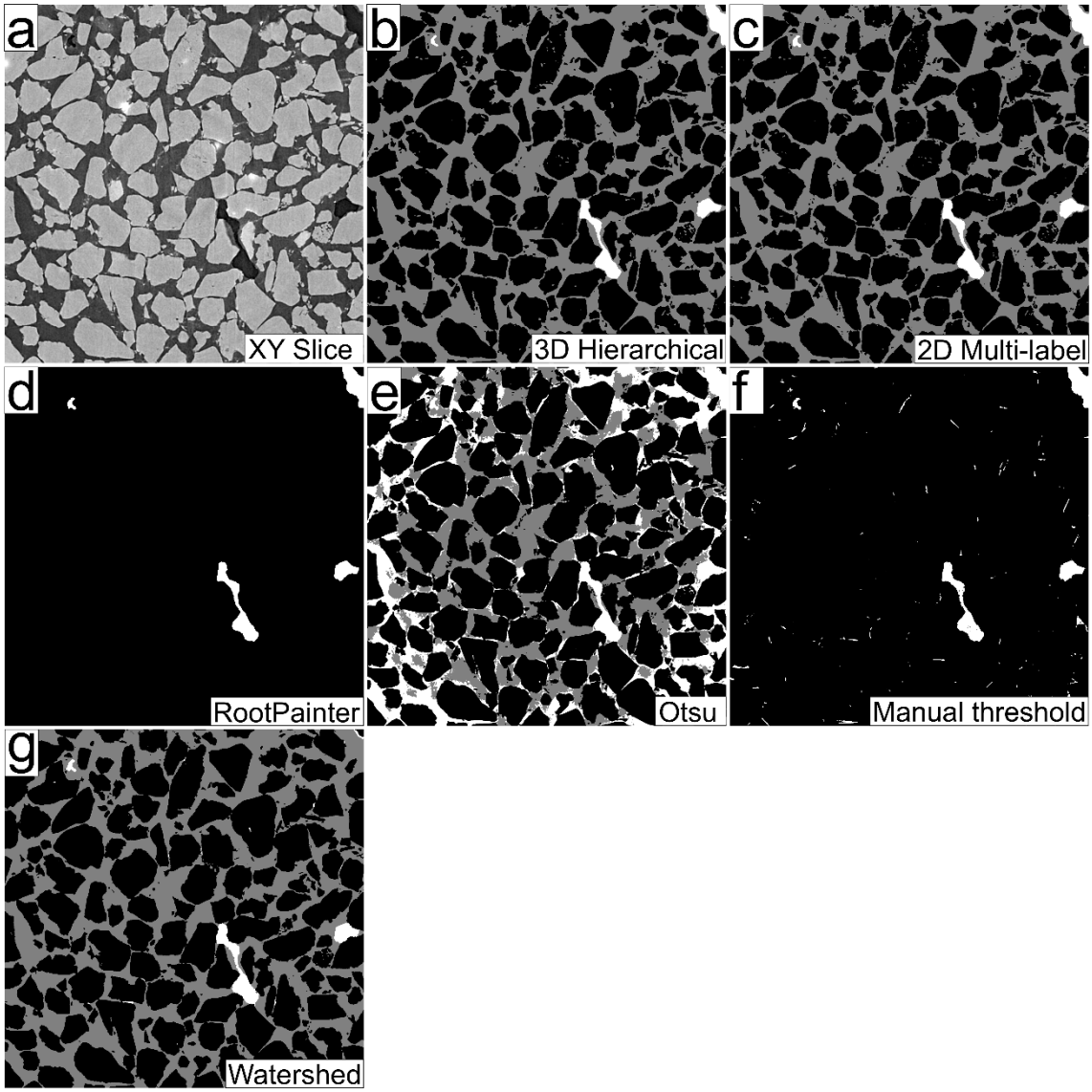

Figure 6 (a) Original XY central slice of data set 89062; (b) Segmented slice using the 3D hierarchical method with the $[384]^3$ training subvolume; (c) Segmented slice using the 2D multi-label single-axis approach; (d) RootPainter segmentation of the CH_4 gas phase; (e) Otsu auto-threshold output; (f) Manual thresholding output; (g) Watershed segmentation. CH_4 gas shown in white, brine in grey and sand in black.

A disadvantage of the use of 2D U-Net segmentation methods that operate solely with XY slices, such as RootPainter and the single-axis 2D multi-label method, is that horizontal stripe artefacts may appear in the vertical (YZ or XY) slices of the segmented volume. This occurs because training and segmentation does not account for feature continuity between slices. Such artefacts are absent in the output of the 3D hierarchical implementation, which is reflected in the "smoothness" of the line showing the per-slice metrics for this approach in Figure 7. These artefacts can be mitigated by predicting segmentation of data slices taken along different axes and subsequently recombining them into a single volume, as done for the multi-axis 2D method, described in Section 2.3.3. This also

improves the algorithm segmentation performance metrics, as shown in Figure 7(b), but at the expense of greater computation times, as presented in Figure 8.

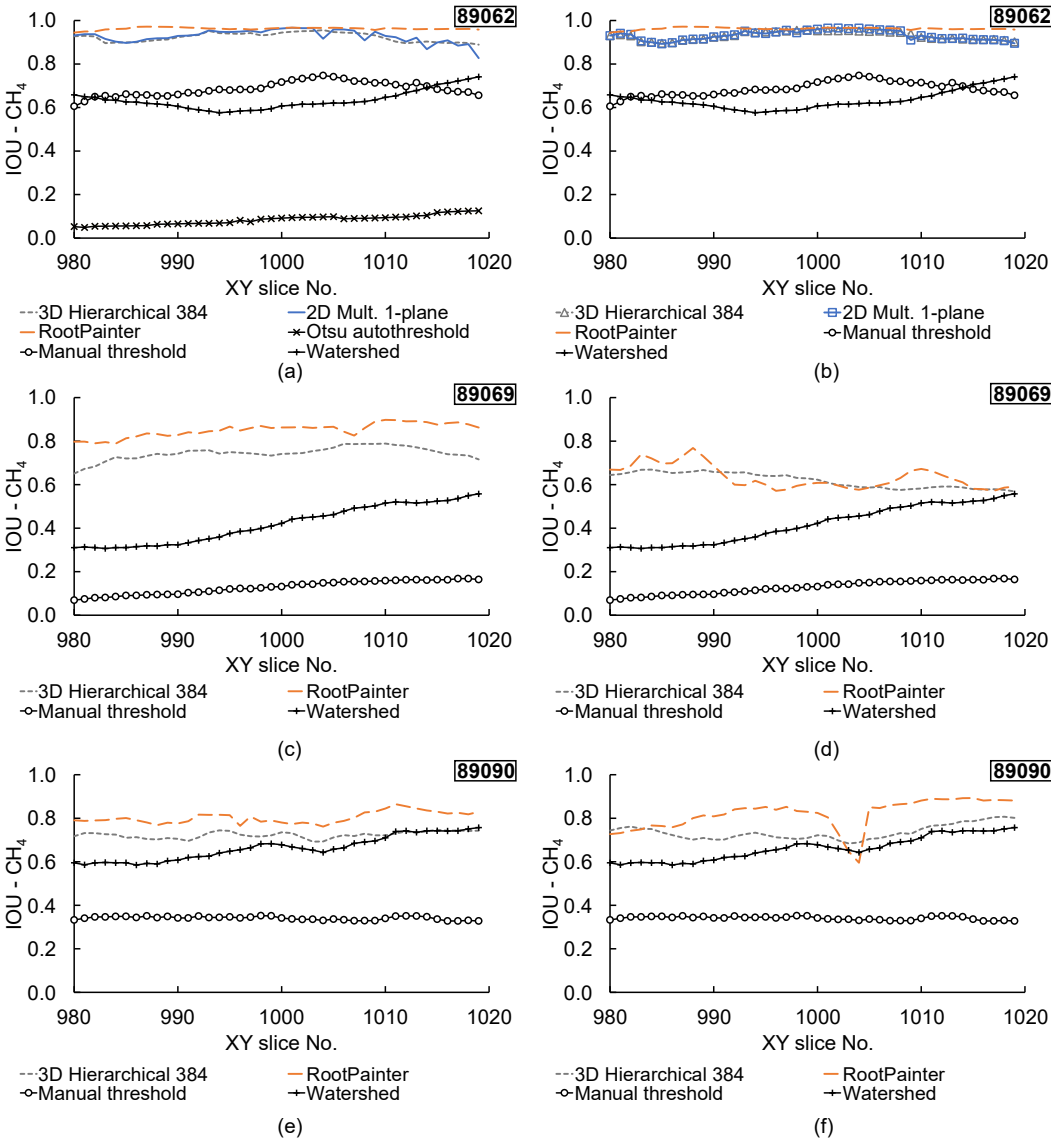

Figure 7 Performance metrics for the segmentation of CH_4 gas on the central 40 XY slices of: (a) 89062 using the 3D hierarchical ($[384]^3$ training sub-volume), the single-axis 2D multi-label and RootPainter U-Nets; (b) 89062 using the 3D hierarchical ($[572]^3$ training sub-volume), the multi-plane 2D multi-label and RootPainter U-Nets; (c) 89069 using the 3D hierarchical ($[384]^3$ training sub-volume) and RootPainter U-Nets; (d) 89069 using the 3D hierarchical ($[384]^3$ training sub-volume) and RootPainter U-Nets trained on data from 89062; (e) 89090 using the 3D hierarchical ($[384]^3$ training sub-volume) and RootPainter U-Nets trained on data from 89062; (f) 89090 using the 3D hierarchical ($[384]^3$ training sub-volume) and RootPainter U-Nets trained on data from 89069. Watershed and thresholding methods shown for reference.

Table 3 Binary 3D hierarchical U-Net training metrics.

Training data source	Labels	Number of training epochs	Final training loss (BCE)	Final validation loss (BCE)	Final validation metric (mean IOU)
89062 - $(384)^3$	CH_4 vs Background	94	0.0313	0.0237	0.935
	Sand vs Background	83	0.0317	0.0332	0.977
89062 - $(572)^3$	CH_4 vs Background	84	0.00551	0.0307	0.918
	Sand vs Background	85	0.0426	0.0290	0.980
89069 - $(384)^3$	CH_4 vs Background	82	0.0178	0.0371	0.759
	Sand vs Background	69	0.0305	0.0471	0.957

Table 4 Single- and multi-axis 2D multi-label U-Net training metrics.

Training data source	Approach	Number of training epochs	Final training loss (CE)	Final validation loss (CE)	Final validation metric (%)
89062 - $(572)^3$	Single-axis (pre-trained)	15	0.020	0.014	99.48
	Multi-axis (pre-trained)	15	0.019	0.012	99.59
Note: % – Percentage of correctly labelled voxels.					

In terms of accuracy, it can be proposed that RootPainter benefits from human intervention via annotative corrections in such way that it can deliver binary segmentations that are marginally superior to the alternative U-Net procedures. However, the alternative U-Net methods are able to (1) segment three material labels with limited user intervention, which results in less user time, and (2) deliver segmentations where horizontal stripe artefacts are largely absent, which can result in higher

quality data visualisation outputs. It is also remarked that U-Net methods are significantly more accurate at segmenting CH_4 gas than the threshold and watershed approaches assessed.

From the user-input and post-processing perspective, all three U-Net methods require the manual annotation of training and validation sub-volumes, which is labour-intensive. RootPainter was able to produce the best segmentation model using only 109 slices for training and validation, including annotative correction slices (included in Alvarez-Borges *et al.* (2021)), but the method can currently only segment one label at a time. Figure 8 also shows that segmenting the CH_4 gas phase using RootPainter requires similar computing resources to segmenting all three labels using the 2D single-axis multi-label approach. On the other hand, while the 3D hierarchical procedure required significantly longer computing times, it produced competitive results and segmented three labels using the $[384]^3$ training and $[256]^3$ validation sub-volumes, which are small compared to the size of the entire 3D image. The off-the-shelf watershed and thresholding methods were not implemented using a graphics processing unit (GPU), in contrast to the U-Net methods, except for the label classification step at the end. Due to this, these standard approaches were executed using solely CPUs on a high-performance computing cluster (HPCC). A qualitative comparison between standard and U-Net methods shows that the watershed technique demands very long computing times (Figure 8). This is partly due to the label classification step, where thousands of feature labels are merged into the desired sand, brine-hydrate and CH_4 gas labels. The thresholding methods proved to be much faster.

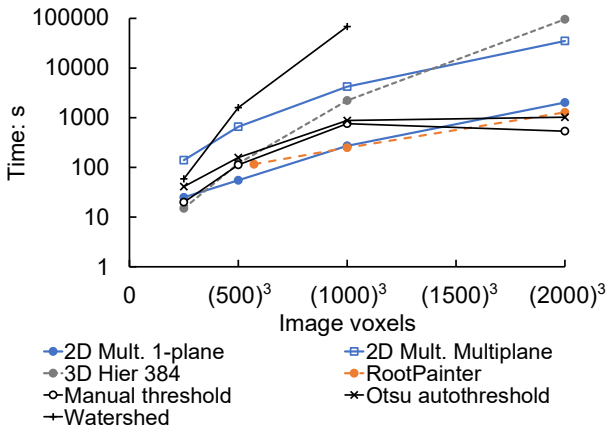

Figure 8 Segmentation time required using an Nvidia Tesla V100® GPU for the U-Net methods and an HPCC of 40×Intel Gold 6242R® CPU @ 3.10GHz for the standard methods. Benchmarking 3D images were extracted from the reconstructed and post-processed scan 89062 and are available from Alvarez-Borges et al. (2021). Excludes time spent on the preparation of training datasets.

3.2. U-Net Performance on data with different greyscale contrast

The trained U-Net models created with all three methods described above we able to predict high-quality segmentations for the intermediate contrast dataset 89062 when trained on subsections of the same dataset. To examine if similar results could be obtained on datasets exhibiting lower greyscale contrast, both 3D hierarchical and RootPainter U-Nets were used to segment 'low' contrast dataset 89069 (Table 1, Figure 1(b)), using $[384]^3$ and $[572]^3$ sub-volumes of the same data for training, respectively. Figure 7(c) presents the performance metrics resulting from this approach, as well as those of watershed and manual thresholding methods applied to the same volume. It may be noted that both U-Net methods return lower metrics than those for the intermediate contrast dataset 89062. The IOU computations show that, on average, 74% and 85% of the voxels predicted to be CH_4 gas were true positives in the 3D hierarchical and RootPainter results, respectively. In comparison, these average values were 92 and 94% for 89062.

Figure 9(a) shows that, for both U-Net methods, the lower performance metrics of the segmentation for 89069 are driven by false positives. However, false positives are over twice as numerous than false negatives in the results for the 3D hierarchical approach, whereas they only surpass false negatives by about 30% in the RootPainter segmentation. For both methods, most false positives correspond to ground truth brine-hydrate voxels incorrectly labelled as CH_4 gas, as depicted in Figure 9(b). This indicates that the reduced grey value differentiation (i.e., contrast) between CH_4 gas and brine-hydrate phases restricted U-Net segmentation accuracy, as anticipated.

Despite this, the U-Net methods significantly out-perform the standard approaches, as shown in Figure 7(c). In fact, performance numbers reveal that watershed and manual thresholding cannot deliver a reliable quantification of the material phases of this dataset.

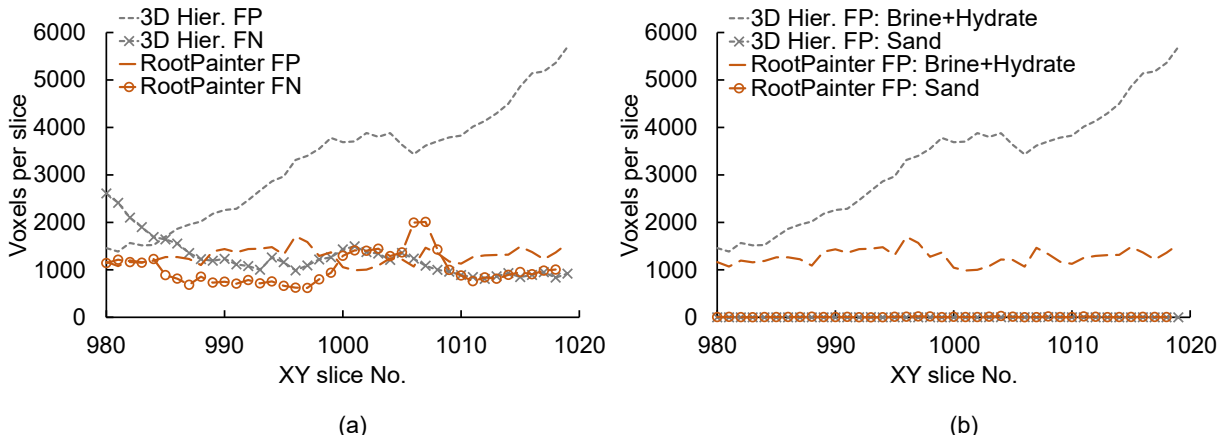

Figure 9 (a) False positive (FP) and false negative (FN) CH_4 gas voxels and (b) FP voxel labels per slice for the central 40 slices of data set 89069 segmented using RootPainter and the 3D hierarchical method (3D Hier).

3.3. U-Net Segmentation Model Generalisation Across Datasets (Model Portability)

To examine U-Net model portability, 'low' and 'high' contrast datasets 89069 and 89090 (Table 1, Figure 1(b)) were segmented using the models produced from training on 'intermediate' contrast dataset 89062. Figure 7(d, e) presents the performance metrics of the resulting segmentations. It may be seen that segmentation accuracy is lowest in the case where the U-Net models trained on mid-contrast dataset 89062 were applied to the low-contrast dataset 89069. IOU values from this process are comparable to those obtained from the thresholding and watershed methods applied to mid-contrast dataset 89062, and thus, quantification from these segmentations may be unreliable. The U-Net model trained on 89062 produced higher accuracy segmentations of high-contrast dataset 89090, comparable to those for the segmentation of low-contrast dataset 89069 using models trained on 89069. Yet, it is evident that U-Net models trained on 89062 perform best when applied to the same 'native' 89062 dataset, as shown by Figure 7(a, e).

As segmentation performance appeared to be higher when U-Net models trained on lower contrast data were used to segment higher contrast data, models trained on low-contrast 89069 images were used to segment high-contrast dataset 89090. Performance metrics are presented in Figure 7(f). This Figure shows an overall improvement in performance metrics compared with segmentations produced with the U-Net models trained on 89062 (Figure 7(e)). However, an instance of localised poor performance for RootPainter can be observed in the profiles of Figure 7(f), which resulted from a cluster of FP pixels on a single slice. This emphasises the limitations of the slice-by-slice (2D) segmentation described in Section 3.1, and denotes a broadly similar pattern of FP-driven model inaccuracy as for the results discussed previously in Section 3.2 (Figure 9).

3.4. Applications and implications

The segmentation of XCT or SRXCT images of soil and rock samples is often carried out to determine parameters such as porosity or liquid/gas saturation, as discussed in Section 1. The varying performances of the U-Net methods used in the present investigation result in differences in the parameters calculated from the segmented images. This is exemplified in Figure 10, which compares porosity and CH_4 gas saturation ratios derived on a slice-by-slice basis from the segmented volumes produced with the 3D hierarchical approach ($[384]^3$ training sub-volume) and RootPainter, which were the procedures that seemed to provide the best results with the least user time. Porosity was calculated as:

$$\text{Porosity } (\%) = \frac{volume\ of\ pores}{total\ volume} \times 100 \tag{2}$$

And CH_4 gas saturation was determined as:

$$\text{CH}_4 \text{ saturation } (\%) = \frac{volume\ of\ \text{CH}_4}{volume\ of\ pores} \times 100 \tag{3}$$

where the volume of CH_4 gas amounts to the total number of CH_4 gas voxels, the volume of pores is the sum of CH_4 gas and brine-hydrate voxels, and the total volume is the total number of voxels in the image multiplied by the voxel volume (0.8125×0.8125×0.8125 μm). For the RootPainter method, the sand phase has been segmented using the same approach used for CH_4 described in Section 2.3.4, but using sand labels and only one quadrant of each annotation slice to produce sparsely annotated training and validation images. Results presented in Figure 10 correspond to two application scenarios, that is:

1. U-Nets trained on sub-volumes of the dataset of interest are then used to segment the entire dataset, shown in Figure 10(a-d). As discussed in Section 3.5, differences in greyscale contrast affect the performance of the resulting segmentation. A training sub-volume needs to be created for each scan.

2. U-Nets trained on sub-volumes of a low-greyscale contrast dataset are then used to segment other 'unknown' datasets of higher greyscale contrast (model portability). This is presented in Figure 10(e-f), corresponding to parameters derived for high-contrast dataset 89090 using segmentations produced from U-Nets trained on sub-volumes of low-contrast dataset 89069. Thus, only one training sub-volume is needed to segment multiple scans.

Porosity and CH_4 saturation calculations derived from manual thresholding and watershed methods are also included in Figure 10. This Figure suggests that, while U-Net models trained on a sub-volume of the same data delivered high segmentation performance metrics for the CH_4 gas phase, the derived parameters deviated from ground truth values to some extent, this being more acute for porosity inferences. In fact, in most cases watershed or thresholding methods delivered more accurate porosity profiles. A comparison between the mean absolute error (MAE) for the porosity and CH_4 saturation calculations along with the mean IOU values for the combined CH_4 gas and sand labels from the three volumes used to generate Figure 10 is shown in Figure 11. This Figure reveals that, while there is a general trend of lower MAE for derived material parameters with higher segmentation accuracy, the correlation exhibits some scatter. Considering that both CH_4 gas saturation and porosity are in part derived using the number of sand voxels and that these are significantly more numerous than pore voxels (CH_4 gas and brine-hydrates), it may be proposed that errors in porosity/CH_4-saturation estimation originate from inaccuracies in the segmentation of the sand phase. This is evidenced in Figure 12 for dataset 89062, which presents (a) IOU metrics for the segmentation of the sand phase and (b) the number of FP and FN voxels. Figure 12(a) reveals that the inaccuracies in the segmentation of the sand phase are relatively small in terms of metrics, which are in fact higher than those of the CH_4 gas phase presented in Figure 7(a). However, Figure 12(b) shows that the number of FP and FN voxels is large compared to the size of the CH_4 gas and brine-hydrate phases, which amount to roughly 3.0×104 and 8.75×105 voxels per slice, respectively. This, in turn, affects

parameters calculated from voxel counts. This denotes that the estimation of soil parameters based on ratios between material phases from segmented images is particularly sensitive to the relative size of said phases. It should be noted, however, that the maximum absolute errors presented in Figure 11 for U-Net-derived parameters (1.40% and 0.26% for porosity and CH_4 gas saturation, respectively) are smaller than those commonly reported for laboratory methods (Matula *et al.*, 2016; Missimer & Lopez, 2018; Péron *et al.*, 2007).

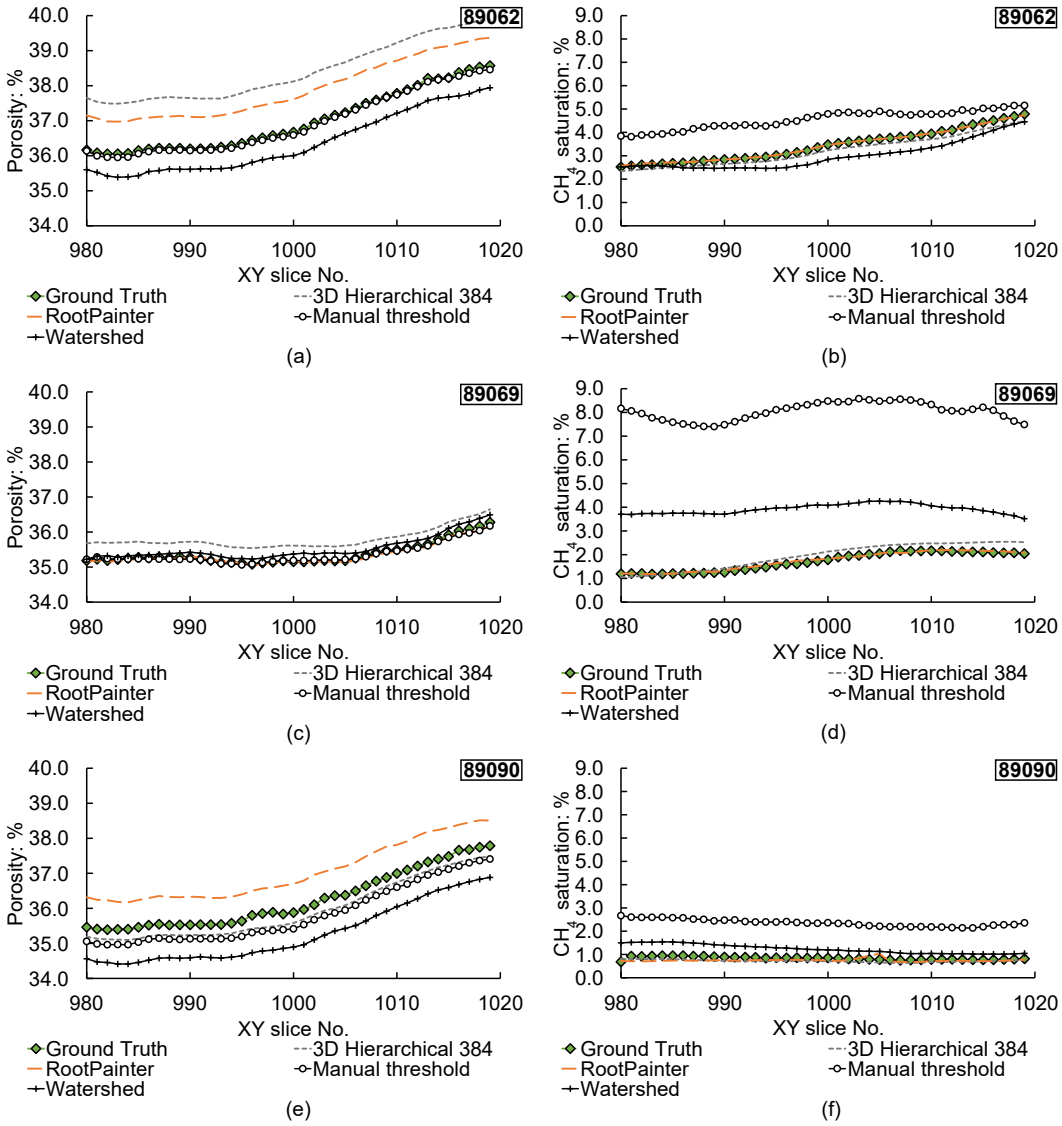

Figure 10 Porosity and CH_4 gas saturation profiles for the central 40 XY slices of data sets 89062 (a, b), 89069 (c, d) and 89090 (e, f) derived using image segmentations obtained from 3D hierarchical and RootPainter U-Net models trained on sub-volumes of 89062 (a, b) and 89069 (c-f). Parameters derived from manual thresholding and watershed segmentation methods also shown.

Figure 12(a) also shows that thresholding and watershed methods are very effective at segmenting abundant, high-contrast, well-defined features like sand, as stated in Section 2.3. Indeed, the use of U-Nets may not be necessary or recommended if only such segmentations are required, as mentioned in Section 2.2.2. However, mainstream methods return unsatisfactory CH_4 gas saturation measurements, particularly for the low-contrast 89069 volume, as shown in Figure 10. This is due to their inability to detect scarce, low-contrast features like CH_4 bubbles, as demonstrated in Section 3.1.

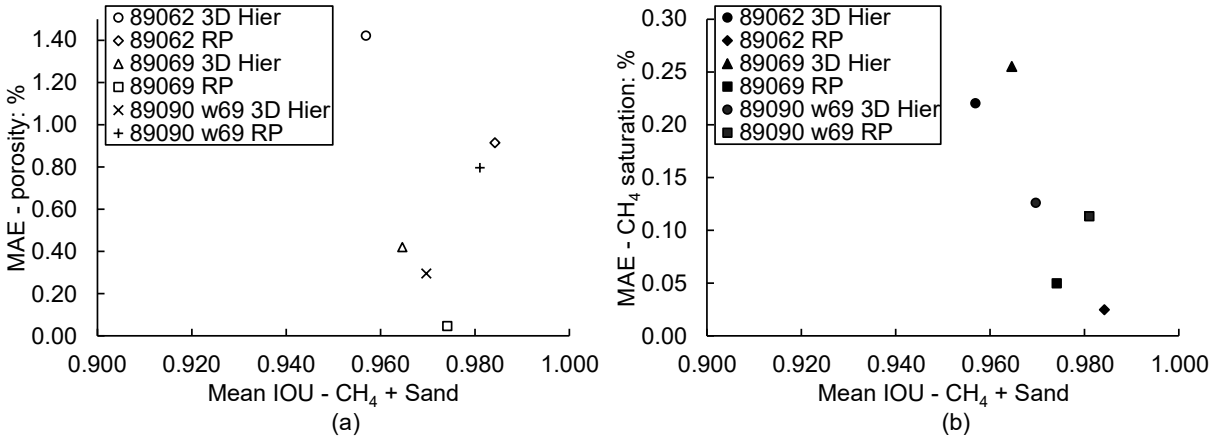

Figure 11 Comparison of mean absolute errors for (a) porosity and (b) CH_4 gas saturation estimations with mean IOU metrics for the segmentations used. w69 denotes the use of a U-Net model trained on a sub-volume of low-contrast dataset 89069; RP refers to RootPainter.

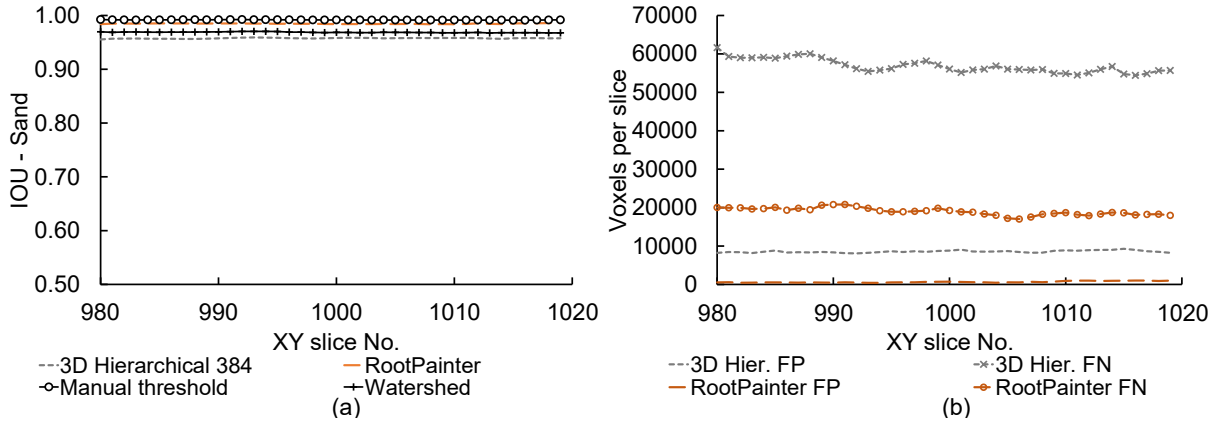

Figure 12 (a) IOU metrics for the segmentation of sand in data set 89062 using the 3D hierarchical method ([384]^3 training sub-volume) and RootPainter, and (b) associated false positive (FP) and false negative (FN) sand voxels per slice of the central 40 XY slices. Metrics for the manual thresholding and watershed methods included in (a) for reference.

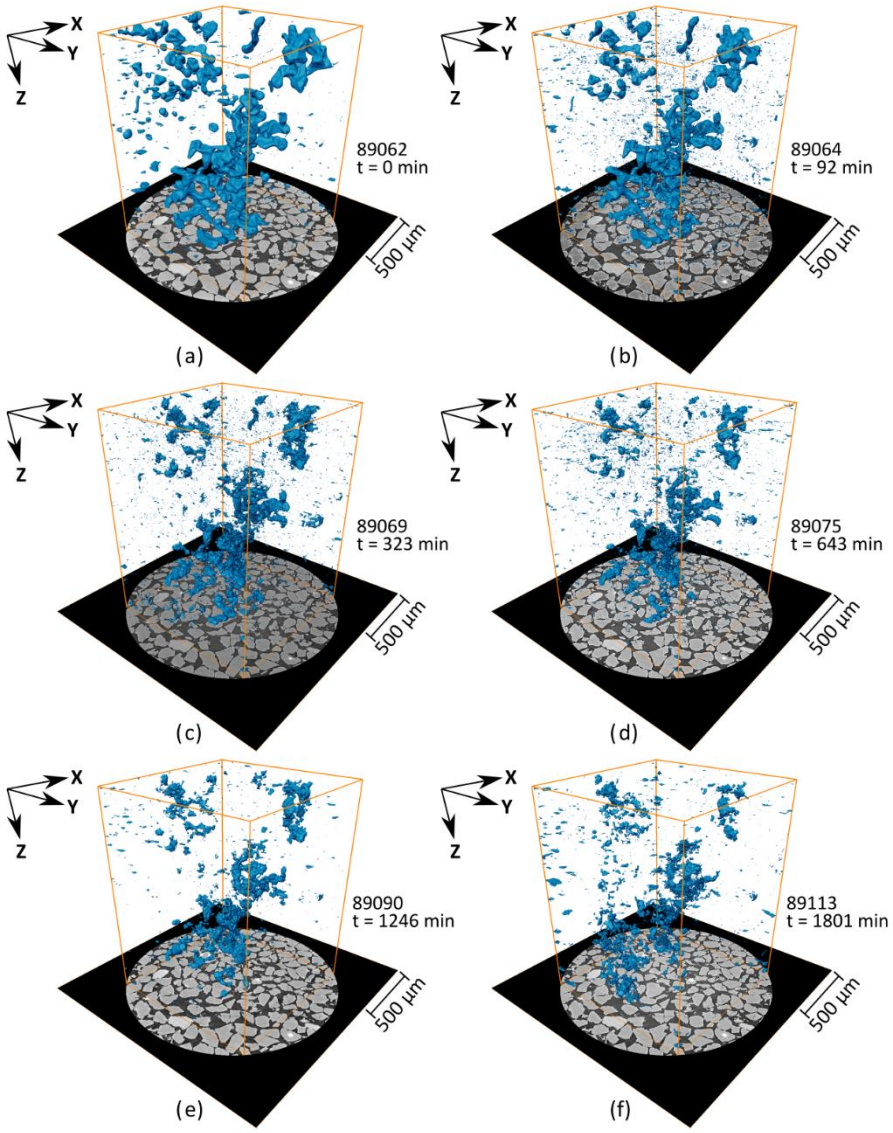

Figure 13 3D views of the CH_4 gas phase segmented using a RootPainter U-Net model trained on the low-contrast 89069 data (t denotes cooling time in minutes after reaching 2°C).

A further application for U-Net segmentations of XCT/SRXCT images of soil and rock is 3D data visualisation, which can then be used to investigate, for instance, CH_4 gas distribution within the pore matrix. Such application can greatly benefit from model portability. To exemplify this, Figure 13 compares 3D views of the CH_4 gas phase produced by segmenting datasets obtained at different stages of hydrate formation using the RootPainter model trained on the low-contrast 89069 sub-volume. Despite the presence of a modest number of segmentation errors in the form of small speckles on some of the images (Figure 13(b-d)), the U-Net model produces sensible 3D representations of the data, and changes in CH_4 gas distribution as it is consumed for hydrate formation can be clearly distinguished. In a further example, a 2D multi-label U-Net, trained using the single-axis approach on the $[572]^3$ volume from scan 89062, has been used to segment a higher-

contrast SRXCT scan from a similar experiment carried out at the Swiss Light Source (SLS) originally reported by Sahoo, Madhusudhan, *et al.* (2018). The post-processing steps described in Section 2.2.2, except cupping correction, were applied to the reconstructed data and a 1554×1554×2000 voxel region was extracted from the centre of the 3D image (data is available from Alvarez-Borges *et al.* (2021)). Results are shown in Figure 14, where it is seen that the model delivers qualitatively accurate 3D views of the distribution of all three material phases, without any additional training or user input.

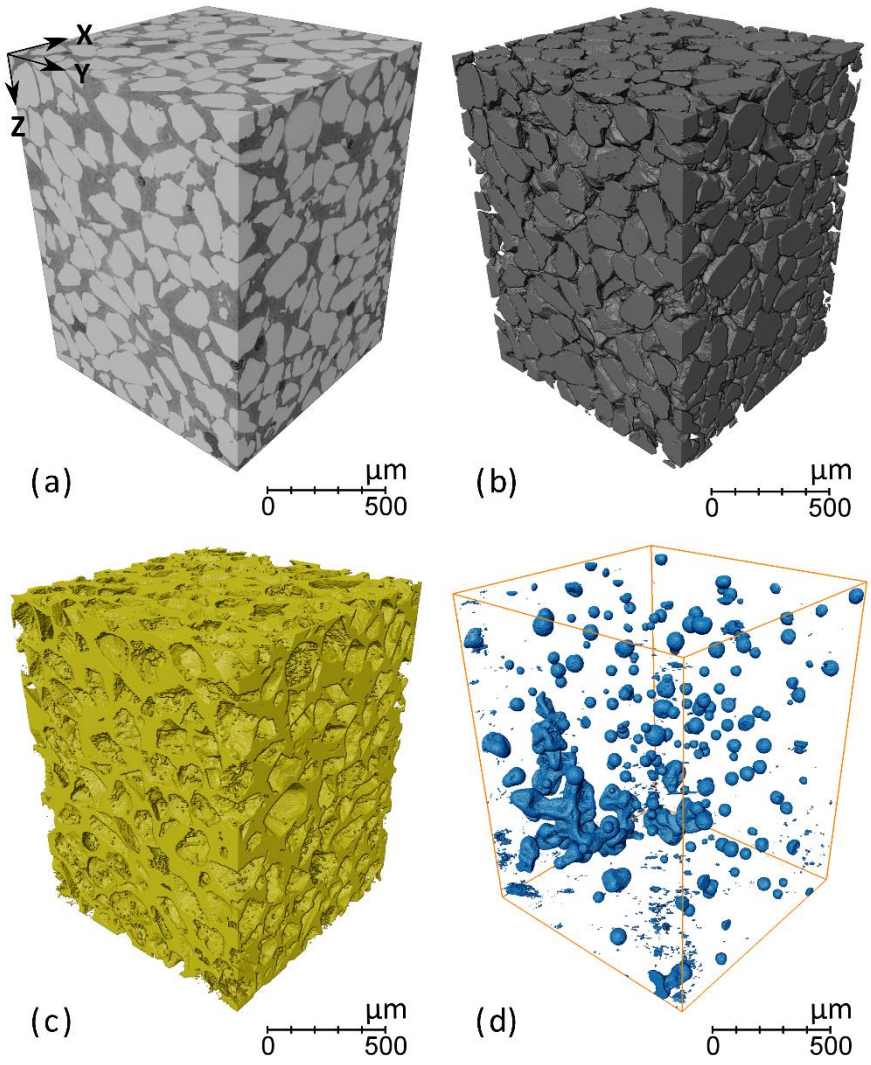

Figure 14 U-Net segmentation of an independent data set from Sahoo et al. (2018a) acquired at SLS, using a 2D multi-label single-axis U-Net model trained on a $[572]^3$ sub-volume of data set 89062: (a) reconstructed SLS volume; (b) sand; (c) brine-hydrate; (d) CH_4 gas.

Both examples demonstrate the capability of U-Net models to segment multiple SRXCT images of CH_4-bearing soil, despite being obtained with different scan set-ups. The U-Net models used only a single $[572]^3$ voxel sub-volume for training and did not require any additional training or user input to segment new images. A key implication is that training of a single U-Net model on a low greyscale contrast dataset could be used to deliver insight on variations in sediment morphology in other datasets. This has valuable applications. For example, segmentations are often required during a short period of time with limited operator input, such as during data acquisition at a synchrotron or other X-ray facility. The availability of pre-trained U-Net models would allow segmentations and sediment morphology/microstructure information to be produced within a short time after acquisition and reconstruction. Pre-trained models could also be used to segment numerous and/or large data sets over shorter timespans with reduced user effort and bias.

4. Conclusions

The application of U-Nets to segment SRXCT images of CH_4-bearing sand has been investigated. The general aim was to determine if these convolutional deep learning networks, trained on a small set of images ($\leq [572]^3$ voxels), were capable of accurately segmenting large SRXCT datasets ($2000\times[1554]^2$ voxels) of different greyscale contrast, with focus on the CH_4 gas phase. Training images were obtained from a hand-annotated subset of the reconstructed SRXCT data. Three U-Net deployment methods were used: 3D hierarchical, 2D multi-label and the RootPainter application. Quantitative comparisons amongst U-Net segmentation outputs, along with mainstream thresholding and watershed methods, were carried out using the IOU metric. Major outcomes of this investigation are presented below.

1. For a given SRXCT data set, the three U-Net deployment methodologies produced models capable of delivering segmented images of the CH_4 gas phase with average IOU metrics of at least 0.74 and up to 0.93. This demonstrated that the U-Net methods used were capable of accurately identifying the CH_4 gas phase using a small number of training images. RootPainter delivered marginally higher IOU metrics than the other methods but suffered from minor horizontal stripping artefacts and required more human intervention and proportionally higher computing time.
2. Greyscale contrast between material phases in the different SRXCT datasets was a significant factor affecting U-Net segmentation accuracy. The lowest segmentation performance metrics corresponded to SRXCT datasets exhibiting the lowest greyscale contrast, while greater segmentation accuracy resulted from the use of higher contrast data.

3. All U-Net segmentations of CH_4 gas outperformed thresholding and watershed methods. However, mainstream methods proved to be more accurate at segmenting abundant, well-defined, and high-contrast features, like sand. U-Net methods are, thus, not recommended for this task.
4. Model portability, i.e., the ability of a U-Net model trained on a subset of one dataset to generalise and produce an accurate segmentation of a different SRXCT dataset, was explored. It was found that models trained on lower-contrast images were able to produce accurate segmentations of higher-contrast data without additional training. In comparison, U-Net models trained on higher-contrast images were found to deliver poor results when used to segment lower-contrast data. Portability was further demonstrated by accurately segmenting independent data from a different synchrotron facility without additional training. This suggests that targeted training on small amounts of 'ground truth' data can produce U-Net segmentation models that can be used for rapid segmentation of a large number of different datasets with additional user input or training.
5. The effect of segmentation accuracy on image-derived material parameters was investigated by calculating porosity and CH_4 gas saturation profiles using U-Net segmentations. A general trend of lower mean absolute error of the derived parameter with greater segmentation accuracy was found, but the correlation exhibited some scatter. Considering that porosity, fluid saturation and other parameters are ratios between material phases, it was proposed that errors in derived parameters are not only linked to segmentation accuracy metrics but to the number of false positive and negative voxel labels of the largest phase relative to the other phases.

Acknowledgements The authors gratefully acknowledge Diamond Light Source for the provision of beamtime (proposal MT16205-1) and are thankful for the support provided by beamline I13 staff. The authors have no conflict of interests to disclose.

References

Alvarez-Borges, F. J., King, O. N. F., Madhusudhan, B. N. & Ahmed, S. I. (2021). *Tomography data of methane-bearing sand used to investigate U-Net segmentation methods [Dataset]*

Atwood, R. C., Bodey, A. J., Price, S. W. T., Basham, M. & Drakopoulos, M. (2015). *Philosophical Transactions of the Royal Society A* **373**, 2369–2393.

Baveye, P. C., Laba, M., Otten, W., Bouckaert, L., Dello Sterpaio, P., Goswami, R. R., Grinev, D., Houston, A., Hu, Y., Liu, J., Mooney, S., Pajor, R., Sleutel, S., Tarquis, A., Wang, W., Wei, Q. & Sezgin, M. (2010). *Geoderma* **157**, 51-63.

Bradski, G. (2000). *Dr. Dobb's Journal of Software tools*.

Brown, W. S. (2016). *Springer Handbook of Ocean Engineering*, edited by M. R. Dhanak & N. I. Xiros, Cham, CH: Springer.

Brunke, O., Brockdorf, K., Drews, S., Müller, B., Donath, T., Herzen, J. & Beckmann, F. (2008). *Proceedings of SPIE 7078, Optical Engineering + Applications*. San Diego, CA, USA: SPIE.

Chauhan, S., Rühaak, W., Anbergen, H., Kabdenov, A., Freise, M., Wille, T. & Sass, I. (2016). *Solid Earth* **7**, 1125-1139.

Chauhan, S., Rühaak, W., Khan, F., Enzmann, F., Mielke, P., Kersten, M. & Sass, I. (2016). *Computers & Geosciences* **86**, 120-128.

Dean, J. F., Middelburg, J. J., Röckmann, T., Aerts, R., Blauw, L. G., Egger, M., Jetten, M. S. M., de Jong, A. E. E., Meisel, O. H., Rasigraf, O., Slomp, C. P., in't Zandt, M. H. & Dolman, A. J. (2018). *Reviews of Geophysics* **56**, 207-250.

Douarre, C., Schielein, R., Frindel, C., Gerth, S. & Rousseau, D. (2018). *Journal of Imaging* **4**.

Fonseca, J., O'Sullivan, C. & Coop, M. R. (2009). *Powders and Grains: Proceedings of the 6th International Conference on Micromechanics of Granular Media*, edited by M. Nakagawa & S. Luding, pp. 223-226. Golden, CO, USA: American Institute of Physics.

Gonda, F., Kaynig, V., Jones, T. R., Haehn, D., Lichtman, J. W., Parag, T. & Pfister, H. (2017). *IEEE 14th International Symposium on Biomedical Imaging*. IEEE.

He, K., Girshick, R. & Dollar, P. (2019). *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4918-4927. Seoul, KOR: Computer Vision Foundation.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778. Las Vegas, NV, USA: IEEE.

Holland, M. & Schultheiss, P. (2014). *Marine and Petroleum Geology* **58**, 168-177.

Howard, J. & Gugger, S. (2020). *Information* **11**.

Hsieh, J. (2015). *Computed Tomography: Principles, Design, Artifacts, and Recent Advances*, 3rd ed. Bellingham, WA, USA: SPIE.

Iassonov, P., Gebrenegus, T. & Tuller, M. (2009). *Water Resources Research* **45**.

IPCC (2013). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, UK and New York City, NY, USA: Intergovernmental Panel on Climate Change.

James, R. H., Bousquet, P., Bussmann, I., Haeckel, M., Kipfer, R., Leifer, I., Niemann, H., Ostrovsky, I., Piskozub, J., Rehder, G., Treude, T., Vielstädte, L. & Greinert, J. (2016). *Limnology and Oceanography* **61**, S283-S299.

Kalender, W. A. (2011). *Computed Tomography. Fundamentals, System Technology, Image Quality, Applications*. Erlangen, DE: Publicis Publishing.

Karabağ, C., Jones, M. L., Peddie, C. J., Weston, A. E., Collinson, L. M. & Reyes-Aldasoro, C. C. (2020). *Plos One* **15**, e0230605.

Karimpouli, S. & Tahmasebi, P. (2019). *Computers & Geosciences* **126**, 142-150.

Kerkar, P. B., Horvat, K., Jones, K. W. & Mahajan, D. (2014). *Geochemistry, Geophysics, Geosystems* **15**, 4759-4768.

King, O. N. F. & Alvarez-Borges, F. J. (2021). *Gas Hydrate Segmentation Using U-Nets. Code Repository*.

Kittler, J. & Illingworth, J. (1985). *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-15**, 652-655.

Kong, D. & Fonseca, J. (2018). *Géotechnique* **68**, 249-261.

Koyuncu, C. F., Arslan, S., Durmaz, I., Cetin-Atalay, R. & Gunduz-Demir, C. (2012). *Plos One* **7**, e48664.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2017). *Communications of the ACM* **60**, 84–90.

Kvenvolden, K. A. (1993). *Reviews of Geophysics* **31**, 173-187.

Lee, K., Zung, J., Li, P., Jain, V. & Seung, H. S. (2017).

Lee, S. U., Yoon Chung, S. & Park, R. H. (1990). *Computer Vision, Graphics, and Image Processing* **52**, 171-190.

Legland, D., Arganda-Carreras, I. & Andrey, P. (2016). *Bioinformatics* **32**, 3532-3534.

Lei, L., Seol, Y. & Jarvis, K. (2018). *Geophysical Research Letters* **45**, 5417-5426.

Loshchilov, I. & Hutter, F. (2019). *Decoupled Weight Decay Regularization [cs, math]*.

Madhusudhan, B. N., Clayton, C. R. I. & Priest, J. A. (2019). *Journal of Geophysical Research: Solid Earth* **124**, 65-75.

Maslin, M., Owen, M., Betts, R., Day, S., Jones, T. D. & Ridgwell, A. (2010). *Philosophical Transactions of the Royal Society A* **368**, 2369-2393.

Matula, S., Báťková, K. & Legese, W. L. (2016). *Sensors* **16**.

Mienert, J. (2009). *Encyclopedia of Ocean Sciences*, edited by J. H. Steele, pp. 790-798. Oxford, UK: Academic Press.

Missimer, T. M. & Lopez, O. M. (2018). *Journal of Geology and Geophysics* **7**, 1000448.

Moridis, G., Collett, T. S., Pooladi-Darvish, M., Hancock, S. H., Santamarina, C., Boswell, R., Kneafsey, T. J., Rutqvist, J., Kowalsky, M. B., Reagan, M. T., Sloan, E. D., Sum, A. & Koh, C. (2011). *SPE Reservoir Evaluation & Engineering* **14**, 76-112.

Otsu, N. (1979). *IEEE Transactions on Systems, Man, and Cybernetics* **9**, 62-66.

Paganin, D., Mayo, S. C., Gureyev, T. E., Miller, P. R. & Wilkins, S. W. (2002). *Journal of Microscopy* **206**, 33-40.

Paris, S., Kornprobst, P., Tumblin, J. & Durand, F. (2009). *Foundations and Trends in Computer Graphics and Vision* **4**, 1-73.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. & Chintala, S. (2019). *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox & R. Garnett, pp. 8024–8035. Red Hook, NY, USA: Curran Associates, Inc.

Pérez-García, F., Sparks, R. & Ourselin, S. (2020). *TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning [Preprint]*.

Péron, H., Hueckel, T. & Laloui, L. (2007). *Geotechnical Testing Journal* **30**, 1-8.

Phan, J., Ruspini, L. C. & Lindseth, F. (2021). *Scientific Reports* **11**, 19123.

Ramachandran, G. N. & Lakshminarayanan, A. V. (1971). *Proceedings of the National Academy of Sciences* **68**, 2236-2240.

Rogowska, J. (2000). *Handbook of Medical Imaging*, edited by I. N. Bankman, pp. 69-85. San Diego: Academic Press.

Ronneberger, O., Fischer, P. & Brox, T. (2015). *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, vol 9351.*, edited by N. Navab, J. Hornegger, W. M. Wells & A. F. Frangi, pp. 234-241. Cham, CH: Springer International Publishing.

Ruppel, C. D. & Kessler, J. D. (2017). *Reviews of Geophysics* **55**, 126-168.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. & Fei-Fei, L. (2015). *International Journal of Computer Vision* **115**, 211-252.

Sahoo, S. K., Madhusudhan, B. N., Marín-Moreno, H., North, L. J., Ahmed, S., Falcon-Suarez, I. H., Minshull, T. A. & Best, A. I. (2018). *Geochemistry, Geophysics, Geosystems* **19**, 4502-4521.

Sahoo, S. K., Marín-Moreno, H., North, L. J., Falcon-Suarez, I., Madhusudhan, B. N., Best, A. I. & Minshull, T. A. (2018). *Journal of Geophysical Research: Solid Earth* **123**, 3377-3390.

Saunois, M., Stavert, A. R., Poulter, B., Bousquet, P., Canadell, J. G., Jackson, R. B., Raymond, P. A., Dlugokencky, E. J., Houweling, S., Patra, P. K., Ciais, P., Arora, V. K., Bastviken, D., Bergamaschi, P., Blake, D. R., Brailsford, G., Bruhwiler, L., Carlson, K. M., Carrol, M., Castaldi, S., Chandra, N.,

Crevoisier, C., Crill, P. M., Covey, K., Curry, C. L., Etiope, G., Frankenberg, C., Gedney, N., Hegglin, M. I., Höglund-Isaksson, L., Hugelius, G., Ishizawa, M., Ito, A., Janssens-Maenhout, G., Jensen, K. M., Joos, F., Kleinen, T., Krummel, P. B., Langenfelds, R. L., Laruelle, G. G., Liu, L., Machida, T., Maksyutov, S., McDonald, K. C., McNorton, J., Miller, P. A., Melton, J. R., Morino, I., Müller, J., Murguia-Flores, F., Naik, V., Niwa, Y., Noce, S., O'Doherty, S., Parker, R. J., Peng, C., Peng, S., Peters, G. P., Prigent, C., Prinn, R., Ramonet, M., Regnier, P., Riley, W. J., Rosentreter, J. A., Segers, A., Simpson, I. J., Shi, H., Smith, S. J., Steele, L. P., Thornton, B. F., Tian, H., Tohjima, Y., Tubiello, F. N., Tsuruta, A., Viovy, N., Voulgarakis, A., Weber, T. S., van Weele, M., van der Werf, G. R., Weiss, R. F., Worthy, D., Wunch, D., Yin, Y., Yoshida, Y., Zhang, W., Zhang, Z., Zhao, Y., Zheng, B., Zhu, Q., Zhu, Q. & Zhuang, Q. (2020). *Earth System Science Data* **12**, 1561-1623.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J. Y., White, D. J., Hartenstein, V., Eliceiri, K., Tomancak, P. & Cardona, A. (2012). *Nature Methods* **9**, 676-682.

Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. (2012). *Nature Methods* **9**, 671-675.

Sell, K., Saenger, E. H., Falenty, A., Chaouachi, M., Haberthür, D., Enzmann, F., Kuhs, W. F. & Kersten, M. (2016). *Solid Earth* **7**, 1243-1258.

Smith, A. G., Han, E., Petersen, J., Olsen, N. A. F., Giese, C., Athmann, M., Dresbøll, D. B. & Thorup-Kristensen, K. (2020). *RootPainter: Deep Learning Segmentation of Biological Images with Corrective Annotation [Preprint]*.

Smith, A. G. & Ørting, S. (2020). *RootPainter 0.2.5*.

Smith, A. G., Petersen, J., Selvan, R. & Rasmussen, C. R. (2020). *Plant Methods* **16**.

Smith, L. N. (2017). *2017 IEEE Winter Conference on Applications of Computer Vision*, pp. 464-472. Santa Rosa, CA, USA: IEEE.

Song, Y., Luo, T., Madhusudhan, B. N., Sun, X., Liu, Y., Kong, X. & Li, Y. (2019). *Journal of Natural Gas Science and Engineering* **72**, 103031.

Titarenko, V., Bradley, R., Martin, C., Withers, P. & Titarenko, S. (2010). *SPIE Optical Engineering + Applications*. Society of Photo-Optical Instrumentation Engineers (SPIE).

Tun, W. M., Poologasundarampillai, G., Bischof, H., Nye, G., King, O. N. F., Basham, M., Tokudome, Y., Lewis, R. M., Johnstone, E. D., Brownbill, P., Darrow, M. & Chernyavsky, I. L. (2020). *A massively multi-scale approach to characterising tissue architecture by synchrotron micro-CT applied to the human placenta [Preprint]*, 2020.2012.2007.411462.

van Aarle, W., Palenstijn, W. J., Cant, J., Janssens, E., Bleichrodt, F., Dabravolski, A., De Beenhouwer, J., Joost Batenburg, K. & Sijbers, J. (2016). *Optics Express* **24**, 25129-25147.

van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., Yu, T. & the scikit-image, c. (2014). *PeerJ* **2**.

Vanneste, M., Sultan, N., Garziglia, S., Forsberg, C. F. & L'Heureux, J.-S. (2014). *Marine Geology* **352**, 183-214.

Varfolomeev, I., Yakimchuk, I. & Safonov, I. (2019). *Computers* **8**.

Vo, N. T., Drakopoulos, M., Atwood, R. C. & Reinhard, C. (2014). *Optics Express* **22**, 19078-19086.

Wadeson, N. & Basham, M. (2016). *Savu: A Python-based, MPI Framework for Simultaneous Processing of Multiple, N-dimensional, Large Tomography Datasets [Preprint]*.

Wadeson, N., Basham, M., Parsons, A., Kazantsev, D., Vo, N. T., Schoonjans, T., Pérez-Juárez, E., mjn19172, Srikanth, N., Nixon, D., Taylor, M., Storm, M., Price, S., Atwood, R. C., Gowling, C., Frost, M., Zdenek, M., Palenstijn, W. J., Badger, T. G., Delicious, M. G., Leinweber, K. & Filik, J. (2019). *DiamondLightSource/Savu: Version 2.4.*

Wolny, A., Cerrone, L., Vijayan, A., Tofanelli, R., Barro, A. V., Louveaux, M., Wenzl, C., Strauss, S., Wilson-Sánchez, D., Lymbouridou, R., Steigleder, S. S., Pape, C., Bailoni, A., Duran-Nebreda, S., Bassel, G. W., Lohmann, J. U., Tsiantis, M., Hamprecht, F. A., Schneitz, K., Maizel, A. & Kreshuk, A. (2020). *eLife* **9**.

Yokohama, T., Nakayama, E., Kuwano, S. & Saito, H. (2011). *Proceedings of the 7th International Conference on Gas Hydrates*, pp. 1830-1833. Edinburgh, UK: Curran Associates Inc.

Zhang, X., Jia, F., Luo, S., Liu, G. & Hu, Q. (2014). *Computer Methods and Programs in Biomedicine* **113**, 894-903.